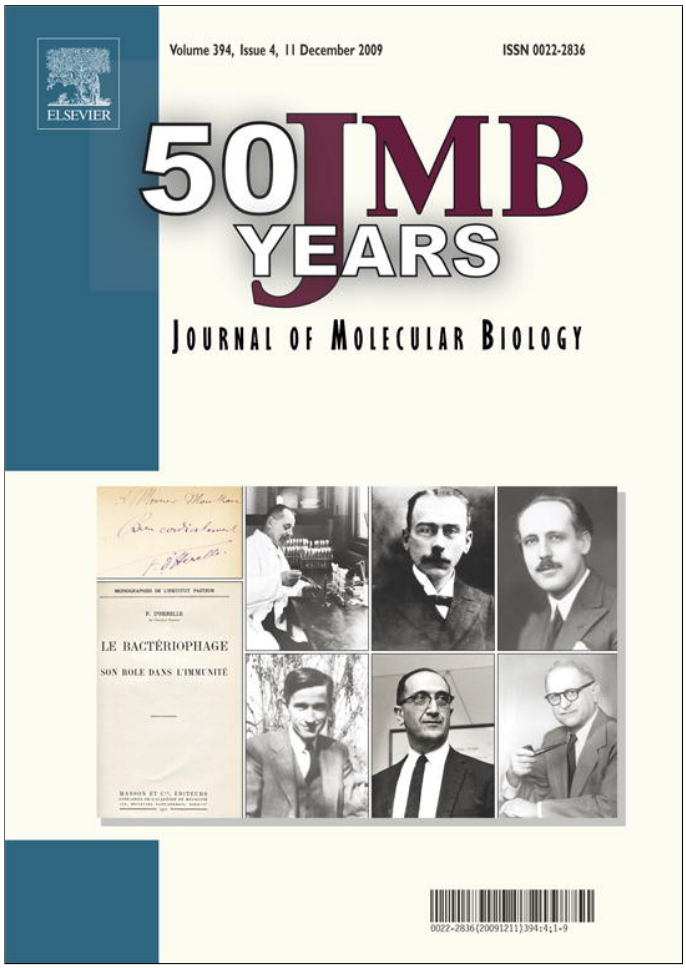

doi:10.1016/j.jmb.2009.09.021 J. Mol. Biol. (2009) 394, 653–680

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Understanding the Differences between Genome Sequences of *Escherichia coli* B Strains REL606 and BL21(DE3) and Comparison of the *E. coli* B and K-12 Genomes

F. William Studier[1]*, Patrick Daegelen[2,3], Richard E. Lenski[4], Sergei Maslov[5] and Jihyun F. Kim[6,7]

[1]*Biology Department, Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973-5000, USA*

[2]*CNRS UMR 8030, Genoscope (CEA), 2 rue Gaston Crémieux, CP 5706, 91000 Evry Cedex, France*

[3]*Inserm, 101 rue de Tolbiac, 75013 Paris, France*

[4]*Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA*

[5]*Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, NY 11973, USA*

[6]*Industrial Biotechnology and Bioenergy Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong, Daejeon 305-806, Korea*

[7]*Functional Genomics Program, University of Science and Technology, Yuseong, Daejeon 305-333, Korea*

Received 17 July 2009; received in revised form 9 September 2009; accepted 10 September 2009
Available online 15 September 2009

Each difference between the genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3) can be interpreted in light of known laboratory manipulations plus a gene conversion between ribosomal RNA operons. Two treatments with 1-methyl-3-nitro-1-nitrosoguanidine in the REL606 lineage produced at least 93 single-base-pair mutations (~90% GC-to-AT transitions) and 3 single-base-pair GC deletions. Two UV treatments in the BL21(DE3) lineage produced only 4 single-base-pair mutations but 16 large deletions. P1 transductions from K-12 into the two B lineages produced 317 single-base-pair differences and 9 insertions or deletions, reflecting differences between B DNA in BL21(DE3) and integrated restriction fragments of K-12 DNA inherited by REL606. Two sites showed selective enrichment of spontaneous mutations. No unselected spontaneous single-base-pair mutations were evident. The genome sequences revealed that a progenitor of REL606 had been misidentified, explaining initially perplexing differences. Limited sequencing of other B strains defined characteristic properties of B and allowed assembly of the inferred genome of the ancestral B of Delbrück and Luria. Comparison of the B and K-12 genomes shows that more than half of the 3793 proteins of their basic genomes are predicted to be identical, although ~310 appear to be functional in either B or K-12 but not in both. The ancestral basic genome appears to have had ~4039 coding sequences occupying ~4.0 Mbp. Repeated horizontal transfer from diverged *Escherichia coli* genomes and homologous recombination may explain the observed variable distribution of single-base-pair differences. Fifteen sites are occupied by phage-related elements, but only six by comparable elements at the same site. More than 50 sites are occupied by IS elements in both B and K, 16 in common, and likely founding IS elements are identified. A signature of widespread cryptic phage P4-type mobile elements was identified. Complex deletions (dense clusters of small deletions and substitutions) apparently removed nonessential genes from ~30 sites in the basic genomes.

*Corresponding author. E-mail address: studier@bnl.gov.

Abbreviations used: SNP, single-base-pair difference; MNNG, 1-methyl-3-nitro-1-nitrosoguanidine; LPS, lipopolysaccharide; indel, insertion or deletion; orf, open reading frame (protein-coding sequence).

Edited by M. Gottesman

Keywords: E. coli B genome; SNP distribution; complex deletions; CP4-type mobile elements; UV deletions

Introduction

The most widely used laboratory strains of *Escherichia coli* have been those derived from strains K-12 and B (referred to here generically as K and B), the result of pioneering work using K for biochemical genetics[1–3] and B for studying virulent bacteriophages[4,5] in the 1940s. The first whole-genome sequence of a K strain, MG1655, was reported in 1997,[6] and its sequence has been compared in detail with that of K strain W3110.[7] Genome sequences of B strains have only recently been determined[8]: REL606 is a strain used for long-term evolution experiments in the laboratory,[9–11] and BL21(DE3) is a strain widely used for production of recombinant proteins under control of T7 RNA polymerase.[12,13] In a companion paper,[14] we trace the ancestry of the *Escherichia coli* B of Delbrück and Luria[4] and the lineages of the two sequenced B strains.

The accompanying paper reporting the genome sequences of the two B strains summarizes many differences between them and provides explanations for some long-known differences between B and K.[8] In the first of two major sections of the present paper, we report detailed comparison of the genome sequences of the two B strains and plausible explanations for every difference between them, buttressed by limited sequencing of other B strains to understand where differences arose. In the second major section of this paper, we analyze in depth and interpret many similarities and differences between the genomes of B and K. Finally, we briefly report DNA sequences of limited regions of the genome of the first *E. coli* strain to be described and isolated, by Escherich in 1885,[15] which was deposited in the UK National Collection of Type Cultures (NCTC) in 1920. We obtained the DNA sequences from the Escherich strain to test whether it might have been the ultimate laboratory progenitor of B. However, we found instead that the Escherich regions we sequenced are, in general, more closely related to K than to B. The three strains are known or surmised to derive from normal commensals of the human gut.[14]

Differences between REL606 and BL21 (DE3) Genomes, and Relationships to Other B Strains

The single circular genomes of REL606 and BL21 (DE3) contain 4,629,812 bp and 4,557,508 bp, respectively, and have a surprisingly large number of differences with a puzzling distribution: 317 of the 426 single-base-pair differences (SNPs) and 9 of the 18 insertions or deletions (indels) of 1–113 bp were found in an ~65-kbp (~1.4%) segment of the genome.[8] The solution to this puzzle and the explanations for the other differences emerged upon further analysis of the genome sequences in light of laboratory manipulations that occurred in the two lineages,[14] together with sequencing certain regions of other B strains, listed in Table 1. The puzzle was solved only when the analyses revealed that a strain had been misidentified in the line of descent to REL606.

Ultimately, we arrived at a satisfactory explanation for each of the differences between the two genome sequences, summarized briefly here and in

Table 1. Bacterial strains

Strain	Single colony	Succession and source	Received
MG1655		Reference sequence	
W3110		Kaiser; Studier	1963
Escherich		1885; NCTC 86 (Lister Institute deposit, 1920)	2007
Bordet	1920	d'Herelle; Bordet 1920; CIP 63.70 (Wollman deposit 1963)	2007
S/6	1946	Bronfenbrenner; Hershey 1946; Doermann 1953; Bolle; Belin	2008
Delbrück		Bronfenbrenner; B 1942; CIP 54.125 (Wollman deposit 1954)	2007
Luria		Bronfenbrenner; B 1942; CIP 103914 (Luria deposit ATCC 1961)	2007
B/r	1946	B Demerec; Witkin 1946; CIP 54.156 (Wollman deposit 1954)	2007
B62	1956	B Luria or Demerec; Bertani; Kaiser 1956; Studier	1963
BB	1958	B; Stent 1958; Bolle; Belin	2008
B40 *sulI*	1961	B; Gorini 1961, 1970; Bolle; Belin	2008
B^E	1968	B; Doermann; R. Epstein 1968; Bolle; Belin	2008
REL606	1961	B; Arber 1961; S. Lederberg 1966; Levin 1972; Lenski	1985
WA251	1959	B; Bc Cohen 1959; Arber 1961 = Bc251; Böhm, 2005	2008
B707	1959	Bc251; Wood 1966; Arber; Studier	1973
B834	1959	B707; Wood 1966; Arber; Studier	1973
B834(DE3)	1959	B834; Studier	1983
BL21	1959	B834; Studier	1978
BL21(DE3)	1959	BL21; Studier	1983

Details about the strains and references are given in the accompanying paper.[14]

All of the B strains in the table, except the Delbrück and Luria strains, are known or almost certain to have originated from a single-colony isolation. The date in the "Single colony" column is the latest date by which the lineage to the current strain must have diverged by single-colony isolation from the B of Delbrück and Luria (or, for Bordet and S/6, from the d'Herelle or Bronfenbrenner strains).

S/6, BB, B40 *sulI*, and B^E from the Geneva collection were received from D. Belin of the Department of Pathology and Immunology of the Faculty of Medicine of the University of Geneva.

The last six strains in the table are in the lineage to BL21(DE3). WA251 from the Arber collection was received from A. Böhm of the Biozentrum of the University of Basel. It is a single-colony isolate made by Böhm in 2005 from an old Arber stab (Arber, personal communication to P.D., 2007).

Table 2. Summary of genome differences between BL21 (DE3) and REL606

	BL21 (DE3)	REL606	REL606 (hidden)
SNPs (426 total)			
rrlH 23S rRNA gene conversion	0	11	
Due to W3110 from P1	0	317	
MNNG-induced in K DNA	0	6	1
MNNG-induced in B DNA	0	84	2
Spontaneous *tsx*, *rpsL*, *fnr*, *btuB*	0	4	
UV-induced	4	0	
1-bp deletions (11 total)			
rrlH 23S rRNA gene conversion	0	2	
Due to W3110 from P1	2	4	
MNNG-induced G deletions	0	3	
Multi-base-pair indels (25 total)			
Due to W3110 from P1	0	3	
Spontaneous	2	3	
UV-induced deletions	17	0	
Unique IS elements			
IS*1*	2	1	
IS*150*	2	2	
Defective prophage in P2 *att*	0	P2*B	
Defective prophage in λ *att*	DE3	λ*B	

Table 2, with details presented in the following sections. Gene conversion in the 23S rRNA gene *rrlH* in the REL606 lineage is responsible for 11 SNPs and two 1-bp deletions. P1 transduction from the K strain W3110 into a progenitor of REL606 is responsible for the numerous SNPs and indels localized in ~1.4% of the genome, which simply represent differences between B and K in this region. The strain misidentification also explains why REL606, but not BL21(DE3), contains a defective P2-like prophage in the P2 *att* site. Two rounds of mutagenesis of progenitors of REL606 by 1-methyl-3-nitro-1-nitrosoguanidine (MNNG) are evidently responsible for 90 SNPs and 3 single-base-pair deletions, as well as 3 hidden SNPs. Two rounds of UV treatment of progenitors of BL21(DE3) were responsible for only four SNPs but for 16 of the 18 large deletions. (The 17th deletion was caused by UV treatment in another strain and imported by P1 transduction; the 18th was spontaneous.) Four spontaneous SNPs were intentionally selected or unintentionally enriched under laboratory conditions and fixed by single-colony isolations. No unselected spontaneous SNPs are apparent.

Gene conversion in *rrlH*

Whole-genome alignment shows that six of the seven pairs of ribosomal RNA operons at equivalent genome positions are perfectly matched between BL21(DE3) and REL606. However, REL606 differs from BL21(DE3) and all of the other B strains listed in Table 1 by 11 SNPs and 2 single-base-pair deletions in a 1368-bp portion of *rrlH*, which specifies the 23S rRNA in the ribosomal RNA operon between *gmhB* and *dkgB*. The differences include isolated single and double SNPs and a cluster of 8 SNPs and 2 single-base-pair deletions in 14 bp (C-CtcttttAT-gGggG in REL606 instead of CACAGAGCAATCTGTG in the other B strains). The REL606 sequence across this region plus flanking sequences has a 2953-bp perfect match in the *rrlE* operon and 2654-bp perfect matches in the *rrlC* and *rrlG* operons. Sequence differences among bacterial ribosomal operons are known to be redistributed by genetic recombination or gene conversion within a genome or between duplicated genomes.[16,17] Such an event must have occurred in *rrlH* and become fixed by single-colony isolation somewhere in the succession of strains from the B of Delbrück and Luria to REL606.

P1 transduction to Mal$^+$ λ^S

E. coli B was known to be Mal$^-$ and resistant to phage λ, and thus in order to work with λ in B, Arber and Lataste-Dorolle isolated Mal$^+$ λ^S derivatives of strains B, B/r, BB, and Bc by P1 transduction from W3110 in 1961.[18] Our sequencing of DNA from an unmodified B strain, B62, shows that the Mal$^-$ λ^R phenotype of ancestral B strains is due to a 5998-bp deletion relative to BL21(DE3), REL606, or K strains (base pairs 4,244,312–4,250,309 of MG1655), which is replaced by an IS*1* insertion element that lacks the customary 8- or 9-bp insertional duplication, consistent with the deletion having been IS*1* mediated. The deletion extends from within *malE* through *yjbI*, eliminating *malK*, *lamB*, and *malM*. PCR analysis confirmed that all of our B strains not descended from a Mal$^+$ λ^S transductant have this deletion but that the K and Escherich strains do not (Table 3). The Mal$^+$ and λ^S phenotypes were later found to be temperature sensitive, probably due to a mutation in the *malT* transcriptional regulator.[19] The sequence of BL21(DE3) shows 14 SNPs in *malT* relative to K, two of which change the amino acid specified: codon 359 of 902 (including the initiation and termination codons) specifies Glu in K but Ala in B, and codon 616 specifies Ile in K but Thr in B. One or both of these changes are presumably responsible for the temperature sensitivity. REL606 has an additional change in codon 766, from His to Asp, evidently caused by MNNG treatment (see below), but we do not know whether it affects the phenotype.

P1 transduction and *EcoB* restriction

W3110 DNA introduced into the B strains by P1 transduction would have been subject to the *EcoB* restriction system, which degrades DNA containing unmethylated TGA(N8)TGCT sequences.[20] To become stably integrated into the host DNA by homologous recombination, the incoming DNA would have to escape degradation and become methylated at *EcoB* recognition sites within it. The 6.0-kbp segment of W3110 DNA that corresponds to the *mal* deletion of B contains one *EcoB* recognition site, which is separated from flanking *EcoB* sites by 13.4 and 14.9 kbp. *EcoB* binds to its recognition sites and translocates DNA toward itself from both sides, cutting the DNA when neighboring enzymes meet.[21] The wide spacing of *EcoB* sites flanking the

Table 3. Differences among B strains and comparison with K and Escherich strains

Strain	*mal*	P2*B[a] bp del	UV	*lon*	*SulA*	*ftsZ*	*fnr* Codon 141	*fnr* aa	*btuB* Codon 58	*btuB* aa	λ*B	*flu*	*fli, dcm*
MG1655	+	*ogr*-D′		+	+	+	cag	Gln	cag	Gln	0	CP4-44	+
W3110	+	*ogr*-D′	R	+	+	+	cag	Gln	cag	Gln	0	CP4-44	+
Escherich	+	0	S	A381V	+	L347Q	cag	Gln	cag	Gln	0		+
Bordet	ISΔ	+	R	+	+	+	tag	stop	tag	stop	+	Phev*B	0
S/6	ISΔ	+	R	+	+	+	cag	Gln	tag	stop	+	Phev*B	0
Delbrück	ISΔ	+	R	IS	+	+	tag	stop	tag	stop	+	Phev*B	0
Luria	ISΔ	+	S	IS	+	+	cag	Gln	tcg	Ser	+	Phev*B	0
B/r	ISΔ	+	R	IS	−7 t >c	+	cag	Gln	tag	stop	+	Phev*B	0
B62	ISΔ	2400	S	IS	+	+	cag	Gln	tcg	Ser	+	0	0
BB	ISΔ	4834	S	IS	+	+	tac	Tyr	ttg	Leu	+	0	0
B40 *suI*[b]	ISΔ	475	R	+	+	+	cag	Gln	tag	stop	+	0	0
B^E	ISΔ	19,776	R	+	+	+	cag	Gln	tcg	Ser	+	0	0
REL606	+	+	R	IS	G31E	+	gag	Glu	ttg	Leu	+	Phev*B	0
WA251[c]	+	0	S	IS	+	+	aag	Lys	6bpΔ	Δ	1.9Δ	0	0
B707	+			IS	+	+	tag	stop	tag	stop	+	short	
B834	+				+	+	tag	stop	ta(g/t)	mix			
B834(DE3)			S		+	+	tag	stop	tat	Tyr	DE3	0	
BL21	+		S		+	+	tag	stop	tag	stop			
BL21(DE3)	+	0	S	IS	+	+	tag	stop	tag	stop	DE3	0	0

All of the B strains in the table, except the Delbrück and Luria strains, are known or almost certain to have originated from a single-colony isolation. No entry in a column means that the strain was not tested.

[a] + refers to the complete insertion in the original B strain; numbers are base pairs deleted from this element in different B strains; *ogr*-D′ refers to the remnant present in the P2 insertion site of K strains.

[b] B40 *suI* also has a deletion of TGAAAA in *fnr* at base pair 1,384,417 of BL21(DE3).

[c] WA251, recovered from an old stab of Bc251 (Table 1), has a 1917-bp deletion in λ*B.

single *EcoB* site in the *lamB* region suggests that the initial cuts might often be far enough apart that a DNA fragment spanning the *lamB* deletion could become recombined into the genome, where its single *EcoB* site should be stable even if not initially methylated (because the methylated flanking sites in the recipient genome would not bind the *EcoB* enzyme complex). This particular scenario may have happened in the P1 transduction that produced Bc251 because its descendant, BL21(DE3), contains a single W3110 fragment smaller than 10.7 kbp that includes the single *EcoB* site covered by the *mal* deletion of B.

By contrast, the REL606 genome contains at least six fragments derived from W3110 DNA totaling ~45–55 kbp interspersed with ~22–26.5 kbp of B DNA across ~71–77 kbp, as determined by comparing the patterns of SNPs and indels between REL606, BL21(DE3), W3110, and B62 (Fig. 1 and Table S1). A single fragment of W3110 DNA covering this entire region would be easily accommodated in the P1 phage particle, which has a capacity of at least 93.6 kbp.[22] The segment of W3110 DNA that covers the 77-kbp region contains nine *EcoB* sites, which are found at comparable positions in B (shown in Fig. 1). Four of the *EcoB* sites are clustered within a 2.6-kbp stretch that was not integrated into the recipient genome, presumably because DNA containing closely spaced *EcoB* sites was cut rapidly and the resulting small fragments were efficiently eliminated by exonucleases. The integrated W3110 fragment that carries the selected *mal* genes is ~21–26.5 kbp and contains two *EcoB* sites separated by 14.9 kbp, at least one of which would have to become methylated for the fragment to become stably integrated. The several other unselected W3110 fragments that also integrated into the genome contained either one *EcoB* site or none. The pattern of integration seems to be consistent with *EcoB* fragmentation of the entering W3110 DNA followed by integration of six of the resulting fragments. The sequenced strain of W3110 has an IS*5* element within one of the fragments acquired by the predecessor to REL606; however, the strain of W3110 used for the P1 transduction probably lacked IS*5* at this position because none is present in REL606 or MG1655 and the sequenced strain of W3110 is known to have insertions not present in other laboratory strains of W3110.[7]

The W3110 DNA fragments introduced by P1 transduction into the genome of the predecessor to REL606, but not BL21(DE3), are responsible for 317 SNPs and 9 indels between these two B strains (Table 2 and Fig. 1; listed individually in Table S1). All of the indels and all but four of the SNPs within this region have been covered by sequences of PCR products of B62 totaling ~48 kbp, and the genome sequence of BL21(DE3) agrees with that of B62 (a B strain not intentionally modified) at every position across this entire region.

Misidentification of a progenitor of REL606

The markedly different patterns of W3110 DNA integrated into the genomes of BL21(DE3) and REL606 show that the two strains are descended from different Mal^+ λ^S transductants, even though the literature says that both descended from Bc251, which is a transductant of Bc, a strain that had been cured of a P2-like defective prophage by Denise

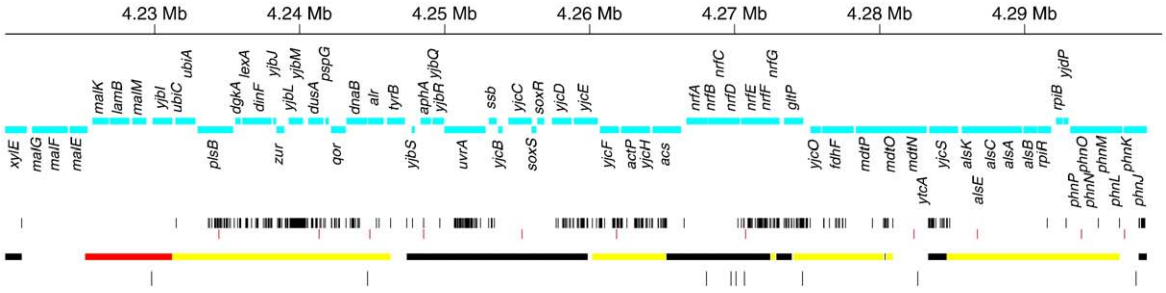

Fig. 1. Pattern of integration of P1-transduced W3110 DNA evident in REL606. The scale gives position in the REL606 genome. The cyan bars show positions of 70 annotated coding sequences in this region, the upper set transcribed clockwise and the lower set counterclockwise. Below the coding sequences, vertical black lines mark positions of all SNPs and indels between the B and W3110 genomes in this region, and vertical red lines in the row below the SNPs and indels mark positions of 11 MNNG-induced mutations in REL606 (all SNPs, indels, and mutations in this region are listed in Table S1). The REL606 genome contains at least six segments derived from W3110 DNA delivered to a progenitor of REL606 by P1 transduction (see the text), and these account for 317 SNPs and 4 indels between REL606 and BL21(DE3). The red bar shows the position of the Mal deletion of B whose replacement was selected for after P1 transduction; yellow bars show the minimum extent of blocks of W3110-derived DNA in the genome; and black bars show the minimum extent of blocks of B DNA into which they integrated, with gaps indicating uncertainty about the exact positions of crossover within identical sequences. The vertical lines at the bottom mark the positions of *EcoB* binding sites in this region of REL606 DNA. W3310 DNA has *EcoB* binding sites at the corresponding positions, and fragmentation of the entering W3110 DNA presumably occurred where *EcoB* enzymes met while remaining bound to their recognition sites and translocating DNA toward themselves from both directions (see the text).

Cohen.[14,23,24] The primary P2 attachment site[25] of BL21(DE3) is unoccupied, as expected for a descendant of Bc251, but that of REL606 is occupied by a 22.1-kbp prophage that has substantial regions of homology to the 33.6-kbp phage P2. We refer to this prophage as P2*B, our designation for the large mobile element characteristic of B in the P2 attachment site. P2*B is clearly defective, having lost three substantial segments of P2 DNA, but it has a complete, highly diverged immunity gene and the necessary excision genes, entirely consistent with its being the prophage cured from B by Cohen.[23] Clearly, REL606 is not descended from Bc or Bc251 and must be descended from one of the other Mal$^+$ λ^S transductants made by Arber and Lataste-Dorolle,[18] namely, those of B, BB, or B/r. The sequence differences among B strains reported in the following sections show that the progenitor of REL606 could not have been BB or B/r and must therefore have been B itself.

P2*B defective prophage

Arber and Lataste-Dorolle[18] observed that λ produced normal plaques on lawns of Mal$^+$ λ^S transductants of Bc and BB, but that it plated significantly less well on transductants of B or B/r. Suspecting that the defective P2-like prophage might interfere with λ growth and that BB might also lack this prophage, they made lysogens of BB and Bc251 with a P2 *Hy dis* phage they obtained from Cohen and found that, indeed, λ also grew poorly on lawns of those lysogens. We therefore tested BB, B/r, and our other strains for the presence of the P2*B defective prophage found in REL606. All of the B strains not descended from Bc, including BB, contained both ends of P2*B. However, PCR sampling across P2*B and selective sequencing of PCR products showed that 4 of these 10 B strains have different internal deletions in P2*B (Table 3): BB has a 4834-bp deletion affecting the P2 tail genes H, I, J, W, and V; B62 has a 2400-bp deletion affecting genes I, J, W, and V; B40 *suI* has a 475-bp deletion affecting I and J; and B^E has a 19,776-bp deletion that leaves only the P2-like *cox*, C, and *int* genes intact. The likely interpretation is that deletion of some or all of these P2 tail genes confers a selective advantage under at least some laboratory conditions, that the deletion in BB had already occurred at the time of the P1 transductions to Mal$^+$ λ^S, and that the Mal$^+$ λ^S transductants of B and B/r did not have a deletion in P2*B. In any case, it is clear that BB is not a progenitor of REL606.

The deletion within P2*B of BB also explains an old observation[26] that *rII* mutants of T-even bacteriophages show the large-plaque, rapid-lysis *r* phenotype on B itself but not on Bc or BB. The part of P2*B that is deleted in BB is apparently necessary for *rII* mutants to give the *r* phenotype.

The Escherich strain and most of the completely sequenced *E. coli* strains have no insertion at the primary P2 *att* site, but K strains are known to contain a fragment of P2 with an active *ogr* gene at this site.[25,27] The P2 fragment in K apparently arose by an aberrant excision between the *att* site at the right end of a P2-type prophage (which could have been P2*B or P2) and a cryptic *att* site in P2 gene D, removing all of the prophage except an intact *ogr* gene and the adjacent C-terminal fragment of gene D. This *ogr*-D′ remnant appears to be unique to K strains among the sequenced strains of *E. coli*. The equivalent 638-bp *ogr*-D′ segments of K, P2*B, and P2 differ by 10, 11, and 10 SNPs, respectively, from the consensus sequence.

UV sensitivity

B was early found to be unusually sensitive to killing by UV irradiation. Witkin, starting with a single colony of B from a culture she received from Demerec, studied the sensitivity of B to radiation and isolated from the culture a mutant with more normal resistance, called B/r.[28,29] The abnormal sensitivity of B to radiation was subsequently found to be due to a deficiency in the ATP-dependent protease specified by *lon*,[30] and the Lon deficiency was caused by an IS*186* insertion in the *lon* promoter, an insertion that is retained in B/r.[31] The increased UV resistance of B/r results from a mutation affecting *sulA*,[32] whose expression is induced by the SOS response to DNA damage.[33] The SulA protein interacts with FtsZ to prevent cell division while the DNA damage is repaired. SulA is normally degraded by Lon, which rapidly removes the protein when repression is reestablished after DNA repair, thereby allowing cell division to resume.[34] The Lon deficiency makes B sensitive to radiation because the SulA induced by DNA damage persists after repair, continuing to inhibit cell division, and the still-growing cells form long filaments that are ultimately lethal. The SulA deficiency in B/r either prevents or relieves the block to cell division so that cells that rapidly repair damage to their DNA are better able to resume growth after an SOS response.

To determine whether B/r could be a progenitor of REL606, we tested our strains for UV sensitivity (loss of colony formation), presence of the *lon* insertion (by PCR), and potential mutations in *sulA* and *ftsZ* (by DNA sequencing) (Table 3). The UV results were surprising because different strains showed different sensitivity over a considerable range, separating the B strains into two broad groups (designated S or R in Table 3). The S group showed sensitivity approximating that reported for B by Witkin, and it includes BL21(DE3) and other descendants of Bc251 as well as the Luria, BB, and B62 strains. The R group showed resistance approximating that of B/r and includes B/r, REL606, and three other strains derived from the B of Delbrück and Luria, as well as the Bordet and S/6 strains derived from progenitors of B. The Escherich strain also fell in the sensitive group, and W3110, as expected, in the resistant group.

UV resistance mutations of REL606 and B/r and identification of B itself as the progenitor of REL606

The difference in UV sensitivity between BL21 (DE3) and REL606 should be evident as a difference in their genome sequences. A mutation of REL606 at base pair 1,037,919 that changed codon 31 of 170 of *sulA* from Gly to Glu is almost certainly responsible because it is the only difference between the two strains in the coding sequences and upstream regions of *lon*, *sulA*, or *ftsZ*. This mutation presumably reduces the ability of SulA to inhibit cell division, thereby increasing the radiation resistance of REL606. This *sulA* mutation was probably caused by the MNNG treatments of progenitors of REL606 because it is a G-to-A transition, and we determined that none of our other B strains has this mutation, nor indeed any difference from the *sulA* coding sequence of K (Table 3).

DNA sequencing found no mutation in the *sulA* coding sequence to account for the UV resistance of B/r[32] but rather an A-to-G mutation at 1,038,047 (in REL606 coordinates), 7 bp ahead of the start of the *sulA* mRNA.[33] This mutation changes the most highly conserved T in the consensus sequence for *E. coli* promoters to C, which should greatly reduce or eliminate *sulA* promoter activity and thereby prevent SOS-induced production of SulA. This *sulA* promoter mutation is unique to B/r and is not present in REL606 or any of the other strains (Table 3).

These results demonstrate that B/r could not have been a progenitor of REL606. Combined with the previously discussed results that rule out Bc or BB as a progenitor, we conclude that the progenitor of REL606 could only have been a Mal^+ λ^S transductant of B itself.

Not all B strains are Lon deficient

The larger-than-expected fraction of B strains relatively resistant to UV irradiation raised the possibility that not all of them carry the IS*186* insertion in the *lon* promoter thought to be typical of B. Indeed, PCR analysis showed that the Bordet, S/6, B40 *suI*, and B^E strains, all of which are relatively resistant to UV, lack this insertion (as do the K and Escherich strains) (Table 3). The absence of the *lon* insertion in the Bordet strain and in the S/6 derivative of the S strain of Hershey is particularly interesting because those strains derive from progenitors of the B strain of Delbrück and Luria rather than from B itself, whereas the B40 *suI* and B^E strains, which also lack the *lon* insertion, are thought to have been independently derived from the B of Delbrück and Luria.[14] This variable presence of the *lon* insertion may have resulted from single-colony isolations from a mixed population in the Bronfenbrenner strain, which gave rise to both the B of Delbrück and Luria and the S strain of Hershey, or from a mixed population in the Delbrück and Luria strain. However, it is also possible that the Delbrück and Luria strain was homogeneous in having the *lon* insertion, and that the B40 *suI* and B^E strains actually derive from strains that Hershey isolated from the Bronfenbrenner strain, which were also referred to as B.[35]

The relative resistance to UV of the Delbrück strain remains unexplained, as this strain has the insertion in the *lon* promoter, which should make it sensitive, and no compensating mutation in *sulA* or *ftsZ* to confer resistance. Sequencing the relevant portions of the Escherich strain found an Ala-to-Val mutation at codon 381 (of 785) in the *lon* coding sequence and a Leu-to-Val change at codon 347 (of 384) in the *ftsZ* coding sequence (Table 3). Perhaps one or both of

these mutations are responsible for the relative UV sensitivity of the Escherich strain.

Evidence for evolution in the laboratory

The *fnr* gene specifies a global regulator of anaerobic growth[36] and the *btuB* gene specifies a transporter of vitamin B_{12} (cobalamin).[37] Both genes have an internal termination codon in BL21 (DE3) that is removed by a SNP in REL606, codon 141 of 251 in *fnr* and codon 58 of 615 in *btuB*. These two SNPs are the only ones in which the corresponding site in B62 did not match either BL21(DE3) or REL606 (Table S1). To try to determine the ancestral sequence, we sequenced PCR products covering these codons for all of our strains (Table 3). The mutational patterns across the different strains are quite variable but suggest that the ancestral Bordet strain[14] had a stop codon at both of these positions, and that both were retained in the B that Delbrück and Luria derived from the coli PC of Bronfenbrenner. However, most of the derivatives of B or the Bronfenbrenner strain lost the termination codon (and presumably regained function) in one or both of these genes. The only strains in Table 3 that retain both termination codons are the Bordet and Delbrück strains from culture collections plus some descendants of Bc251, including BL21(DE3). At least four independent mutations occurred at each TAG codon: to CAG, TAC, GAG, or AAG in *fnr* and to TCG, TTG, TAT, or a 6-bp deletion in *btuB*.

The Fnr regulatory protein is not functional during aerobic growth[36] and the absence of the BtuB transporter is completely compensated by the presence of 1 μM or more vitamin B_{12} in the growth medium.[38] Quite possibly, both genes were irrelevant for aerobic growth on the rich, nutrient agar slants typically used by early workers for the maintenance of strains. It also seems possible that the initial *btuB* mutation was enriched by selection because BtuB is the receptor for at least one phage, BF23,[39] and the culture from which an early progenitor of B was isolated was apparently contaminated with phage.[14] Perhaps laboratory conditions more prevalent later, such as anaerobic conditions in sealed agar stabs and growth media with lower B_{12} concentrations (possibly tryptone broth), provided a selective advantage for mutants that reacquired the *fnr* and *btuB* functions. Evidently, such mutants were independently enriched in several laboratory populations, and individual mutations then became fixed when single-colony isolates were made.

Improved SeMet labeling with BL21(DE3)

The *btuB* mutations at codon 58 are interesting in another regard. Strain B834 has a defective *metE* homocysteine methylase and requires for growth either methionine or B_{12}, which activates the *metH* homocysteine methylase.[40] Our culture of B834 turned out to be a mixed population containing either TAG or TAT at codon 58 of the *btuB* high-affinity B_{12} transporter. Sequencing showed about an equal mixture of T and G at the third position, and that culture gave rise to both B834(DE3), which has TAT, and BL21, which retains the internal TAG termination codon. As little as ~0.75 nM B_{12} was found to support growth of B834(DE3) in the absence of methionine,[40] demonstrating that *btuB* is functional. SeMet labeling of target proteins expressed in B834(DE3) in defined medium was stimulated by 100 nM B_{12}, apparently by activating the B_{12}-requiring *metH* homocysteine methylase to regenerate SeMet from the selenohomocysteine produced by methylation reactions that consume SeMet. However, such stimulation was less apparent in BL21(DE3). This reduced stimulation can now be understood as an inability to transport enough B_{12} to provide complete activation of the *metH* enzyme, and can presumably be compensated for simply by adding at least 1 μM B_{12} to the defined labeling medium.[38]

SNPs generated by manipulations of progenitors to REL606

REL606 is the clonal descendant of a strain that resulted from four rounds of selection or screening applied sequentially by Seymour Lederberg to the Mal^+ λ^S transductant of B that he mistakenly thought was Bc251.[14,24] The individual mutations responsible for the selected phenotypes as well as the spectrum of mutations produced by two treatments with MNNG are revealed by comparing the genome sequences of REL606 and BL21(DE3) with the sequenced K strains, which are typically ~99% identical with B in comparable regions, and with sequenced PCR products of B62 and other B strains. These comparisons established that BL21(DE3) almost certainly carries the base-pair characteristic of B at all but 4 of the 426 SNPs between REL606 and BL21(DE3) (Table 2).

A spontaneous mutation of REL606 selected for resistance to phage T6 is evident as a change in codon 258 of 295 in the *tsx* gene from Trp to the termination codon TAG (bp 399,663). A subsequent spontaneous mutation selected for streptomycin resistance is evident as a change in codon 43 of 125 in the *rpsL* gene from Lys to Thr (bp 3,402,330), a mutation known to confer streptomycin resistance.[41] This double mutant was then subjected to two rounds of treatment with MNNG. The first treatment generated a mutant lacking *EcoB* restriction and modification, which must be due to one or more of the only three mutations in the *hsd* genes[8]: codon 219 of 475 in *hsdS* changed from Thr to Ile (base pair 4,558,782), codon 165 of 530 in *hsdM* changed from Pro to Ser (base pair 4,560,531), and codon 753 of 1171 in *hsdR* changed from Pro to Ser (base pair 4,562,480). The second MNNG treatment generated a mutant unable to grow on L-arabinose, due to a mutation that changed codon 92 of 501 in *araA* from Gly to Asp (bp 70,867). A second mutation that changed codon 256 from Ala to Thr (bp

Table 4. Multi-base-pair indels between BL21(DE3) and REL606

Start	End	Length (bp)	Affected genes or region	Note
Deletions in BL21(DE3) relative to REL606				
REL606 base-pair positions				
Spontaneous, including ompT deletion obtained along with selected IS*1* insertion in *hsdS*:				
555,164	573,218	18,055	DLP12 end + nearby *Rhs*	ompT and ~20 other genes
2,191,465	2,191,469	5	In *mglB*	Deletion of GACAT
UV-induced Gal deletion of Bc258 introduced by P1 transduction:				
769,048	786,294	17,247	16 genes, *galM–ybhJ*	Upstream of λ attachment site
Deletions presumed mostly to have been caused by UV treatments to obtain Bc and Met$^-$ mutant:				
285,494	288,517	3024	In *eaeH*	K-12 strains have IS*3* inserted in *eaeH*
539,637	539,910	274	In DLP12 lambdoid prophage	
551,981	552,851	871	In DLP12 lambdoid prophage	
883,824	887,193	3370	In Rybb*B Fels2-type phage	
1,412,788	1,414,288	1501	In Rac lambdoid prophage	In *recE*
1,420,627	1,423,170	2544	In Rac lambdoid prophage	
1,498,430	1,500,852	2423	In *RhsE* element	
1,605,434	1,615,503	10,070	In Qin lambdoid prophage	Deletes an IS*3*
1,620,090	1,621,452	1363	In Qin lambdoid prophage	
1,626,289	1,628,383	2095	In Qin lambdoid prophage	
1,740,673	1,741,526	854	*sufA–sufB*	In-frame *sufA–sufB* protein fusion
2,006,448	2,007,794	1347	In CP4-44 mobile element	
2,998,430	3,010,189	11,760	In Phev*B CP4-type element	Deletes *flu* gene
3,384,876	3,385,811	936	In *gspD*	Internal in-frame deletion
3,396,797	3,397,429	633	In *chiA*	Internal in-frame deletion
3,697,185	3,700,445	3261	In *RhsA* element	
P2∗B defective prophage in REL606				
2,100,308	2,122,453	22,146	Between *yegQ* and *yegR*	Prophage excised>Bc>BL21(DE3)
Deletions in REL606 relative to BL21(DE3)				
BL21(DE3) base-pair positions				
2,042,391	2,042,402	12	*yegE*	Deletion of GCACCCAACTCG
3,810,659	3,810,663	5	*gidB*	Deletion at CTCTTCT; replication error?
4,166,333	4,166,362	30	*yjbL* in-frame deletion	From W3110 by P1 transduction
4,166,933	4,166,938	6	*yjbM* in-frame deletion	From W3110 by P1 transduction
Insertions in BL21(DE3) relative to REL606				
BL21(DE3) base-pair positions				
748,411	791,335	42,925	(+) DE3 prophage in λ *att*	Replaced 12,090-bp λ*B element at base pairs 787,879–799,968 in REL606
2,068,532	2,069,299	768	(+) IS*1* in *gatR*	Insertion duplication: TGAAATCG
2,819,532	2,820,974	1443	(+) IS*150* in *kduD*	Insertion duplication: ACC
3,822,340	3,820,898	1443	(−) IS*150* in *rbsD*	Insertion duplication: GGT
4,486,413	4,485,646	768	(−) IS*1* in *hsdS*	Insertion duplication: CAAAATTG
Insertions in REL606 relative to BL21(DE3)				
REL606 base-pair positions				
2,128,600	2,129,367	768	(+) IS*1* in *gatZ*	Insertion duplication: GTTTCGACG
2,775,877	2,774,435	1443	(−) IS*150* between *hokX–cysH*	Insertion duplication: GGT
3,080,238	3,080,319	82	(+) 82-bp repeat in *hybO*	82-bp perfect tandem duplication
3,893,554	3,894,996	1443	(+) IS*150* between *kup–rbsD*	Insertion duplication: GAC
4,274,899	4,275,011	113	113-bp repeats	From W3110 by P1 transduction

70,376) is evidently irrelevant, since spontaneous Ara$^+$ mutants[9] reverted only the mutation in codon 92.

Spectrum of mutations induced by MNNG

The two MNNG treatments of progenitors of REL606 apparently generated 90 SNPs, 3 deletions of a single GC base pair, and at least 3 "hidden" SNPs, which are single-base changes not apparent as SNPs between REL606 and BL21(DE3) (Table 2). The hidden SNPs are a mutation at 4,261,847 that eliminated a B/K SNP in the acquired W3110 DNA and mutations at 2,101,939 and 2,121,481 in regions of the P2∗B prophage, not present in BL21(DE3), which were sequenced in other B strains. The three separate deletions of a single GC base pair have no counterparts in BL21(DE3) and thus seem likely to have been caused by MNNG, which is known to attack G. Some 107 kbp of the REL606 genome are not present in the genome of BL21(DE3) (Table 4), and thus a few additional MNNG-induced mutations might be present in REL606. Previous work has established that GC-to-AT transitions are by far the most frequent class of mutation caused by MNNG,[42] and such transitions represent 89% (85/96) of the mutations presumably induced by MNNG in progenitors of REL606 (Table 5). The remaining eight SNPs include three AT-to-GC transitions, four GC-to-TA transversions, and one GC-to-CG transversion.

These data suggest that each MNNG treatment produced roughly 50 single-base-pair mutations in the cells selected for further work. The 96 mutations identified as having been induced by MNNG changed 87 different codons in 77 different coding sequences, and 55 of them (63%) changed the specified amino acid (including 4 changes to a termination codon). Some of the mutations clearly

Table 5. Mutations caused by MNNG or UV or that arose spontaneously in the two B lineages

	Type of single-base-pair change						Deletions (bp)					
	G>A C>T 1	A>G T>C 2	G>C C>G 3	G>T C>A 4	A>C T>G 5	A>T T>A 6	1	5–12	>200	IS	82-bp repeat	Total
MNNG	85	3	1	4			3					96
UV	3			1		1			17			22
Spontaneous	1	1	1	1	3	2		4	1	7	1	22

inactivated or changed the function of the protein they affected, but the cumulative effect is unlikely to have substantially impaired growth under standard laboratory conditions because otherwise these strains would not have been chosen for further work. We do not know what fraction of the population was killed by the MNNG treatments given, but it seems evident—and impressive—that some fraction of *E. coli* cells was able to survive ~50 simultaneous random single-base mutations with little impairment of growth.

Deletions are the predominant mutation caused by UV irradiation

To our knowledge, none of the progenitors of REL606 was subjected intentionally to UV irradiation. However, progenitors of BL21(DE3) were subjected twice to UV treatments that left ~0.1% survivors,[14] first in curing B of P2*B[23] and subsequently in isolating a Met⁻ derivative of Bc251 by UV mutagenesis and penicillin selection.[43] Thus, the whole-genome signature of UV-induced mutations is revealed by comparison of the genome sequences of BL21(DE3) and REL606. Surprisingly, BL21(DE3) contains only four SNPs that can be attributed to the two UV treatments but 16 deletions, ranging from 274 bp to 11,760 bp and totaling 46.3 kbp (Tables 2 and 4). The deletions appear to have been caused by the UV treatments because no comparable deletions are found in REL606. Another UV-induced SNP and deletion, both within *metE*, are responsible for the Met⁻ phenotype of B707 and B834 but were removed from B834 by P1 transduction from B62 in creating BL21.[14] The *metE* deletion in B834 removed base pairs 3,905,237–3,906,014 of BL21(DE3). The Gal⁻ phenotype of BL21(DE3), generated by UV treatment of Bc251 and introduced by P1 transduction to create B707,[14] was also due to a deletion (Table 4). Nine of the 18 deletions thought to have been caused by UV arose between 2-bp crossover repeats containing only A or T or both, and eight others contained such a dimer within crossover regions of 3–7 bp. The last arose without a crossover but had TA immediately adjacent to one end. T is known to be readily modified by UV.

Sequencing all of our B strains at the positions of these four UV-induced SNPs showed that they are present only in strains in the BL21(DE3) lineage and pinpointed when each mutation must have been generated. The C-to-A mutation at 2,075,015 apparently arose in the UV treatment used to cure B of P2*B to produce Bc, whereas the T-to-A at 2,007,877 and the G-to-A at 3,996,130 apparently arose in the UV treatment to isolate the Met⁻ derivative of Bc251. The C-to-T mutation at 862,722 is close enough to the UV-induced *gal* deletion to have been transferred along with it by P1 transduction, or it too could have been generated in producing the Met⁻ strain. The UV-induced SNP in *metE* (along with the deletion) is a G-to-A mutation at 3,904,891 that produced an in-frame TAG termination codon. All five of the single-base mutations caused by UV could have involved pyrimidine dimer formation, as each pyrimidine of the base pair that was mutated was in a run of 2–5 consecutive pyrimidines. Three of the UV-induced mutations were C-to-T transitions and the other two were T-to-A and C-to-A transversions (Table 5).

Although some deletions might have accumulated by chance or by selection due to some unknown difference in maintenance and storage in the lineages of BL21(DE3) and REL606 over the half century that separates them from their last common ancestor, the two rounds of UV killing of progenitors of BL21(DE3) almost certainly caused most or all of them. Most of these deletions lie within defective prophages or genes known to be dispensable (Table 4). Overall, these results indicate that UV treatment to a survival of about 0.1% produced, on average, about 8 deletions greater than ~270 bp and two SNPs in the surviving cells.

Spectrum and frequency of spontaneous SNPs

The 10 spontaneous SNPs in *tsx*, *rpsL*, *fnr*, *btuB*, and *sulA* already discussed are distributed rather uniformly across the six possible types of single-base-pair mutations, in contrast to the predominance of G-to-A transitions caused by MNNG (Table 5). All of these spontaneous SNPs were isolated by selective enrichment. Interestingly, the genome sequences of BL21(DE3) and REL606 show no definitive evidence of other SNPs that cannot be readily explained by gene conversion, mutagenic treatment, selective effect, or acquisition of DNA by P1 transduction.

BL21(DE3) and REL606 diverged from the B of Delbrück and Luria of 1942 by the time the first derivatives in those lineages were made by Cohen in 1959 and by Arber in 1961, and the two sequenced strains are separated from each other by at least 18 single-colony isolations.[14] The apparent absence of unselected spontaneous SNPs between these two

strains can be used to estimate an upper bound on the rate of SNP accumulation under typical laboratory conditions. Expansion from a single cell to a typical working stock requires ~34 binary divisions (2^{34}, $\sim 1.7\times10^{10}$). The actual number of generations between the single cells that gave rise to the succession of strains that led to the two sequenced strains was almost certainly higher in some cases because strains maintained for years by serial transfer between slants or by storage in sealed agar tubes before the next derivative was made would have undergone additional divisions before the next single colony was isolated. Neither lineage is known to have been maintained as lyophilized or frozen stocks before 1978, and strains in the BL21 (DE3) lineage are known to have been maintained mostly in sealed agar stabs and probably briefly by serial transfer on agar slants between ~1961 and 1978.[14] We use the conservative estimate of 34 generations between single-colony isolates to estimate an upper bound on the rate of accumulation of unselected spontaneous SNPs. Given a genome of ~4.6 Mbp and 612 cell generations (18×34) separating the two sequenced strains, a single unselected spontaneous SNP would correspond to a point estimate of $\sim 3.6\times10^{-10}$ SNPs per base pair per generation. Because no unselected spontaneous SNPs were apparent, and because the actual separation between the two sequenced strains was almost certainly more than 612 generations, the actual rate of SNP accumulation is probably lower. Our genome-wide estimate falls within the range of previous estimates based on experimental and comparative studies involving many fewer genes.[44–47]

The absence of unselected spontaneous SNPs between BL21(DE3) and REL606 does not imply that certain SNPs might not routinely be present in populations at considerably higher frequencies than implied by this low mutation rate. Certain mutations may accumulate to higher levels as a consequence of higher-than-typical mutation rate of some sequence motifs or by selection, such as the *fnr* and *btuB* mutations discussed in previous sections.

Spontaneous deletions, duplication, and transpositions

In contrast to the absence of unselected spontaneous SNPs since their last common ancestor, REL606 has acquired deletions of 5 bp and 12 bp plus an 82-bp perfect tandem duplication, and BL21(DE3) has acquired spontaneous deletions of 5 bp and 18,055 bp (Table 4). The small deletions are likely due to replication errors, but the mechanism that produced the 82-bp tandem duplication is unknown. None of these mutations of REL606 is found in the other B strains listed in Table 1, and those of BL21(DE3) are present only in its own lineage, as demonstrated by sequencing PCR products from those regions. The 5-bp deletion in the BL21 lineage arose in B834 and produced a mixed culture, as indicated by a mixed sequence at this site in DNA amplified from B834. The BL21 derivative of B834 contains this deletion, whereas the B834(DE3) derivative does not. Curiously, the B834 culture also was a mixed population of *btuB* alleles, as discussed in the earlier section on SeMet labeling. The deletion and the *btuB* mutation apparently arose independently because they assorted differently to BL21 and B834(DE3). The 18-kbp deletion in BL21 (DE3) apparently arose in the B707 cell that was selected as a spontaneous mutant lacking restriction modification activity to isolate B834,[48] which was due to IS*1* insertion into the *hsdS* gene.[8] That event deleted ~20 genes at one end of and flanking the DLP12 defective prophage, including the *ompT* outer membrane protease that is sometimes a problem for purifying intact proteins expressed in *E. coli*.[49]

After their lineages diverged, one transposition by IS*1* and two by IS*150* occurred in the REL606 lineage, and two transpositions by IS*1* (one of them the insertion into the *hsdS* gene)[8] and two by IS*150* occurred in the BL21(DE3) lineage (Table 4). Progenitors in the REL606 lineage are not available for testing, but the lengths of PCR products from strains in the BL21(DE3) lineage showed that the IS*1* transposition into the *gatR* gene first appeared as a mixed population in B834, from which it passed to BL21 but not to B834(DE3). The IS*150* transposition into *kduD* appears to be present in the entire population of BL21 but is not apparent in B834 or B834(DE3), suggesting that it, too, arose in B834 and was present in the cell that was transduced to isolate BL21. None of the transpositions in the REL606 lineage and only one of the transpositions in the BL21(DE3) lineage is present in the other B strains.

The fourth transposition in BL21(DE3), IS*150* into *rbsD*, is present in all of the strains in the BL21(DE3) lineage and also in the Bordet, Delbrück, Luria, and BB strains. No such insertion is apparent in the B/r strain, an early single-colony isolate from the B of Delbrück and Luria, or in the S/6, B40 *suI*, or B^E strains that may have derived directly from the Bronfenbrenner strain (as discussed in the earlier section on Lon deficiency). A mixed population with regard to this IS*150* insertion is apparent in B62, another early derivative of B. Sequencing showed that all of these IS*150* insertions were at the same site in *rbsD*. A mixed population was also evident in B834, but was due to a deletion that removed most of the IS*150* insertion and a portion of *rbsD*. A possible explanation for the distribution of this insertion among B strains comes from evolutionary studies on REL606.[50] One of the IS*150* insertions unique to REL606 is between the *kup* and *rbsD* genes, where it has given rise to deletions that extend into the *rbsD* gene and inactivate the ability to transport ribose. Such mutants have increased fitness under the conditions of the evolution experiment and increase in the population. The IS*150* insertion directly into the *rbsD* gene seen in other B strains might also improve fitness under some laboratory conditions, and it seems likely that this transposition has arisen independently more than once in the B strains.

λ*B insertion

BL21(DE3) is a deliberately constructed lysogen of phage DE3, a λ derivative with phage 21 immunity and an inducible gene for T7 RNA polymerase inserted into its *int* gene.[12] Disruption of the *int* gene prevents normal integration or excision, and thus the *int* function must be supplied from another source to obtain DE3 lysogens. It was a surprise to find that REL606 has a 12,090-bp insertion in the λ attachment site.[8] PCR analysis showed that all of our B strains that are not DE3 lysogens carry this insertion, which we refer to as λ*B, our designation for a large mobile element characteristic of B in the λ attachment site (Table 3). Strain WA251 (recovered from an old stab of Bc251, Table 1) has a 1.9-kbp deletion in λ*B, but the other B strains, including BL21, appear to contain the same λ*B insertion as REL606. The DE3 prophage of BL21(DE3) displaced the resident λ*B of BL21. Comparison of λ*B and phage λ reveals 98.6% sequence identity over a 1811-bp segment that constitutes the insertion module of λ, comprising the *att*P region, the *int* and *xis* genes, the *int* promoter, and ~250 bp of unannotated sequence upstream of the *int* promoter (27,534–29,344 in λ DNA).[51,52] Insertion at the λ *att* site split the module at the integration crossover site, placing the *int* and *xis* genes and their upstream elements at the left end of λ*B and leaving the portion of *att*P distal to the crossover site at the right end. The internal 10.2 kbp of λ*B has little or no similarity to λ DNA, but as much as one-third of it shows similarity in stretches of 0.3–2.2 kbp to sequences annotated mostly as conserved hypothetical or phage-related proteins. The λ*B element could represent either the remnant of a prophage whose active prototype has yet to be sequenced or some other type of mobile element that has appropriated the λ insertion module.

Genome sequence of B-DL, the *E. coli* B of Delbrück and Luria

The genome sequences of REL606 and BL21(DE3), together with sequences of selected regions of the genomes of other B strains, make it possible to infer with confidence the genome sequence of the *E. coli* B of Delbrück and Luria, or at least the genome of the major component if their culture was not a single-colony isolate. This genome sequence, which we name B-DL, was reconstructed from that of REL606 by the following 12 steps in sequence: (1) correct 82 MNNG-induced and 4 spontaneous SNPs outside of the region that was contributed by P1 transduction from W3110 to a progenitor of REL606 (including restoration of the internal termination codons in *fnr* and *btuB*; (2) replace the entire region contributed by W3110 with sequence from BL21(DE3); (3) remove the IS*150* located between *kup* and *rbsD* and the insertional duplication GAC; (4) insert CTTCT to restore the 5-bp deletion in *gidB*; (5) remove the 82-bp duplication in *hybO*; (6) remove the IS*150* between *hokX* and *cysH* and the insertional duplication GGT; (7) add G at base pair 2,618,781 to make AGT and restore MNNG deletion in IS*1*; (8) remove IS*1* in *gatZ* and the insertional duplication GTTTCGACG; (9) add G at base pair 940,075 to make AGT and restore MNNG deletion in *clpA*; (10) add G to make GGG and restore MNNG deletion between *dapB* and *carA*; (11) incorporate the Mal deletion of B62 and associated IS*1* from strain B62; and (12) correct 11 SNPs and 2 single-base-pair deletions in *rrlH*. The possibility remains that one or more hidden SNPs from the two MNNG treatments of REL606 may be present in the parts of the ~101 kbp of genome deleted from BL21(DE3) that have not yet been sequenced in other B strains. If any such hidden SNPs are present, they are likely to lie within large mobile elements, since most of the DNA deleted came from them (Table 4).

Comparison of the B and K Genomes

Many biochemical and genetic phenomena have been studied in both B and K since the 1940s, when both came into wide laboratory use, and these two strains are known to have many similarities as well as some characteristic differences. Restrictions in the ability to grow phages interchangeably or to exchange genetic information between the two strains were prominent in the discovery and elucidation of host modification and restriction of DNA. Before and after this barrier to DNA transfer was discovered and removed, many hybrid strains were constructed, demonstrating that the two strains have similar genomes. As DNA sequencing became feasible, many genes were also shown to be very similar at the DNA level. Completion of the genome sequences of the two B strains showed that they have the same gene order as K strain MG1655[6,7] (GenBank U00096) without genomic inversions, and that the B and K genomes align with >99% bp identity over ~92% of their genomes.[8]

Here we compare the B and K genome sequences in greater detail, point out prominent similarities and differences, and seek to provide plausible explanations for how the differences arose. Because BL21 (DE3) has many deletions and the B-DL sequence was not assembled until after these analyses were done, REL606 usually represents B in the comparisons, correcting as appropriate for the presence of K DNA due to P1 transduction and the spontaneous or chemically induced mutations. The genome sequences of REL606 (GenBank CP000819) and BL21(DE3) (GenBank CP001509) were compared with that of MG1655[6,53] (GenBank NC_000913). Although direct comparison of the B and K sequences is sufficient for understanding many of their differences, other differences can be resolved only by comparison with other completely sequenced genomes of *E. coli* and *Shigella* strains. Eight *E. coli* strains and five *Shigella* strains available in GenBank at the start of our analyses were used to help understand differences between B and K: *E. coli* strains E24377A (CP000800), HS (CP000802), APEC O1 (CP000468),

UT189 (CP000243), CFT073 (AE014075), 536 (CP000247), O157:H7 Sakai (NC_002695), and O157:H7 EDL933 (NC_002655), and *Shigella* strains *S. boydii* Sb227 (NC_007613), *S. sonnei* Ss046 (NC_007384), *S. flexneri* str. 2a 301 (AE005674), *S. flexneri* str. 2a 2457T (AE014073), and *S. flexneri* 5 str. 8401 (NC_008258).

Protein-coding sequences

In comparing the protein-coding sequences of B and K, we concentrated on what we refer to as the basic genome, that which remains after eliminating the large mobile elements or IS elements, which are discussed in later sections. Starting with the complete set of coding sequences annotated for MG1655 in EcoCyc, we eliminated 292 that lie within large mobile elements or in *Rhs* elements; another 65 annotated as *ins* or *int*, which identify transposases associated with IS elements or integrases associated with other mobile elements; and the annotated *yfjZ* gene of K, which seems likely to be an unmatched accidental open reading frame (orf) in the remnant of a complex deletion. The MNNG-induced and intentionally selected spontaneous mutations present in coding sequences of REL606 were corrected to the base pair typical of B, and the coding sequence of BL21(DE3) DNA was used where REL606 contains K DNA acquired by P1 transduction. The 6.0 kbp that replaced the *malE–yjbI* deletion characteristic of B strains is unavoidably K DNA. The resulting 3943 annotated coding sequences in the basic genome of K were aligned individually with the comparable coding sequence at the same genome location in B, if present, and differences were classified according to their predicted effect on the proteins specified (Table 6). A total of 3793 coding sequences at equivalent positions in the two basic genomes could be matched, either entirely or to segments interrupted by IS elements or duplications or to portions remaining after partial deletion. Another 11 coding sequences of K and 12 of B in the two regions of the genome specifying proteins for producing O antigen and core lipopolysaccharide (LPS) did not match any coding sequence at a comparable position in the other genome.[8] Finally, 139 coding sequences annotated in the basic genome of K were completely deleted in B, whereas 93 annotated in the basic genome of B were completely deleted in K. Summing the matched and deleted genes and adding 14 to allow for unmatched O-antigen and LPS core genes gives a total of ~4039 coding sequences occupying ~3.96 Mbp in the basic genome of the common ancestor, of which 97.6% remain evident in K and 96.5% in B.

Most of the proteins specified by the basic genomes of B and K are highly similar (Table 6), and more than half of them are identical: 1253 (33.0%) of the 3793 matched pairs of coding sequences are identical and 878 (23.1%) have only SNPs that do not change the amino acids specified. Proteins of exactly the same length and greater than 91% amino acid identity, which are likely to be functionally equivalent, are specified by 3620 (95.4%) matched pairs. Another 24 (0.6%) matched genes of comparable lengths are more highly diverged (less than 90% bp identity) but seem likely to be functional in one or both strains, and 54 (1.4%) have seemingly minor differences such as small in-frame deletions or changes near their ends that may have little or no effect on function. (In the absence of experimental information, genes were categorized as functional if they seem likely to produce a protein containing more than ~92% of the amino acids of the presumed intact protein.) The remaining 95 (2.5%) matched pairs appeared to have one or both genes defective because of partial deletions, duplications, frameshifts, insertions of IS elements, or SNPs that created an internal termination codon or eliminated an initiation codon. Of these, 17 appear to be defective in both strains; 49 appear likely to be functional in K but defective in B; and 29 appear to be functional in B but defective in K (only some of which are annotated as pseudogenes in K). In addition, one gene each of B and K in the unmatched O-antigen region is interrupted by an IS element. Including the 232 genes that appear to have been completely deleted from the basic genomes of B or K, a total of 329 different proteins presumably specified by the ancestral basic genome (8.1% of the presumed ancestral complement) appear to be defective in B, K, or both. Specific differences in functional capacities of B and K will be discussed in succeeding sections.

Distribution of SNPs

The distribution of SNPs among the set of 3620 matched and presumably functional coding sequences of the basic genomes is highly variable across the genome. As expected if these coding sequences are under selection for functional proteins, silent codon changes are in substantial excess over codon changes that affect the amino acid specified, with an observed ratio of 5.9 rather than the ratio of less than one in three expected for random SNPs. These 3620 coding sequences contain a total of 26,320 SNPs affecting 25,755 codons, an average of 7.3 SNPs or 7.1 affected codons per coding sequence (average length, 971 bp). It is striking that 1253 genes have no SNPs at all when the average number of SNPs per gene is 7.3. A Poisson distribution for randomly distributed SNPs would predict that only about 3 genes should be free of SNPs. Furthermore, the decrease in number of coding sequences with increasing numbers of SNPs per gene is a complex function (Fig. 2). An initial rapid decrease in number of coding sequences as the number of SNPs per coding sequence increases from 0 to 1 to 2 (1253 to 445 to 232 coding sequences) is roughly consistent with a Poisson distribution of the ~0.5 SNP per coding sequence for these genes (dashed line in the figure). However, this rapid initial decrease changes to a more gradual, exponential decrease in the 1540 coding sequences having 4 or more SNPs.

Table 6. Comparison of annotated protein-coding sequences in the basic genomes of MG1655 and REL606

	Coding sequences in the basic genomes[a]						Base pairs of basic genome occupied			
	MG1655		REL606							
	+	0	+	0	Ancestral	%	MG1655	REL606	Ancestral	%
Genes within the basic genomes										
Matched genes of equal length										
Both intact, >92% bp identity and >91% amino acid identity	3620		3620		3620	89.6	3,516,097	3,516,097	3,516,097	88.9
Both defective		13		13	13	0.3	10,185	10,185	10,185	0.3
Internal stop codon in MG1655	2	4	6		6	0.1	12,273	12,273	12,273	0.3
Internal stop codon or loss of start codon in REL606	11		4	7	11	0.3	14,397	14,397	14,397	0.4
Matched genes of unequal length										
Small difference at C-term, ancestral sequence uncertain	6		6		6	0.1	4377	4338	4386	0.1
Both defective		3		3	3	0.1	5549	5024	6266	0.2
MG1655 gene altered	16	19	35		35	0.9	43,894	50,742	50,742	1.3
REL606 gene altered	53		21	32	53	1.3	68,553	56,301	68,553	1.7
Matched genes interrupted by IS elements										
MG1655 genes interrupted	3	7	9	1	10	0.2	13,884	14,409	14,409	0.4
REL606 genes interrupted	12		2	10	12	0.3	11,448	11,328	11,448	0.3
Matched genes with less than 90% bp identity	24		24		24	0.6	22,770	22,593	22,770	0.6
Total of matched genes between B and K	3747	46	3727	66	3793	92.3	3,723,427	3,717,687	3,731,526	94.0
Unmatched genes in O-antigen region (IS*5* in K; IS*1* in B)	6	1	9	1	10	0.2	7017	10,455	10,455	0.3
Unmatched genes in the LPS core region	4		2		4	0.1	3972	2010	3972	0.1
Genes deleted										
Genes in MG1655 deleted from REL606	139				139	3.4	121,322		121,322	3.1
Genes in REL606 deleted from MG1655			93		93	2.3		89,109	89,109	2.3
Total genes within the basic genome	3896	47	3831	67	4039	100	3,855,738	3,819,261	3,956,384	100
Percent of combined basic genome coding sequences	97.6	96.5					97.5	96.5		
% of total genome occupied by basic genome coding sequences							83.1	82.5		
Total genome length							4,639,675	4,629,812		
MG1655 genes in mobile elements										
MG1655 genes in large mobile elements	292					7.2				
Other MG1655 genes annotated as ins or int	65					1.6				
Unmatched *yjfZ*, accidental orf? remnant of complex deletion	1									
Annotated genes of MG1655 in this analysis	4301									

+ columns tabulate matched genes likely to be functional, with potential altered product >92% of matched gene.

0 columns tabulate matched genes likely to be defective.

[a] Separately annotated fragments of the same gene are combined and counted once.

Genes having no SNPs are distributed around the genome at more than 545 different sites containing one or more contiguous SNP-free coding sequences. The longest stretch of contiguous, completely identical coding sequences contains 17 genes extending for 19,261 bp (~0.4% of the genome) and has no SNPs even in the noncoding intervals. Coding sequences with substantially more than the average number (or density) of SNPs are also widely distributed around the genome.

The observed distribution of SNPs seems to be consistent with the inference that horizontal transfer of segments of DNA from diverged genomes accounts for a far larger fraction of SNPs than do point mutations in *E. coli* strains.[54,55] On the genome-wide scale, repeated homologous genetic recombination since the B and K lineages separated from their last common ancestor may be responsible for most or all of the observed regions of high SNP density. The recombined segments would have been acquired from genomes in the natural *E. coli* population having varying degrees of divergence. The SNP-poor regions may represent undisturbed, vertically inherited regions of the basic genome that are gradually accumulating point mutations. Such horizontal transfer in the basic genomes of B and K is apparent in the gene clusters for O antigen and the LPS core, which are known to be under strong selection, and in a region coding for DNA restriction enzymes, which may also have conferred a selective advantage.[8] However, most genes in the basic genomes of B and K are functionally equivalent, and it is possible that many or even most variants acquired

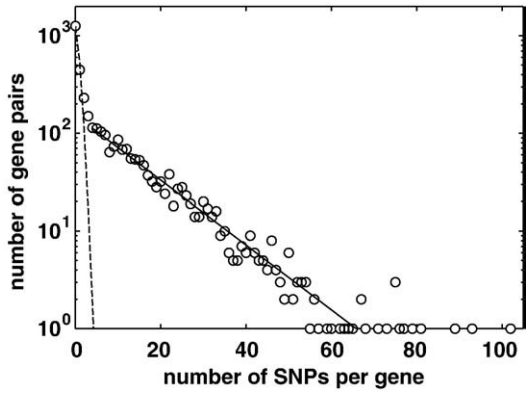

Fig. 2. Distribution of 3620 matched pairs of apparently functional coding sequences of identical length in the basic genomes of B and K, according to the number of gene pairs having a given number of SNPs. The dashed line is a Poisson distribution fit to 0, 1, and 2 SNPs per gene, and the continuous line is the exponential best fit to 4 or more SNPs per gene. The three genes with the largest number of SNPs are *alsA*, with 89 in 1656 bp; *hsdM*, with 93 in 1590 bp; and *yehI*, with 102 in 3633 bp.

by horizontal transfer conferred little or no selective advantage. Nevertheless, it seems plausible that horizontal transfer and fixation of both positively selected and essentially neutral DNA fragments may have generated the observed genome-wide SNP distributions.

Ribosomal RNAs and tRNAs

The sequences and locations in the genome that specify ribosomal RNAs and tRNAs are essentially the same for B and K, except that B is lacking the *ileY* tRNA gene due to an IS*1*-mediated deletion[8] that also removed the right end of the Ssra∗B mobile element (described in the later section on large mobile elements). Alignments of the seven rRNA operons of B and K revealed seven different sites where multi-base-pair polymorphisms extend across 8–37 bp; four of these polymorphisms consist solely of SNPs and three contain small indels as well, with 2–5 different variants of each polymorphism represented among the 14 rRNA operons. Two of these multi-base-pair polymorphisms lie within the sequence specifying 16S RNA, one between the sequences specifying 16S and 23S rRNAs, and four within the sequence specifying 23S RNA. The apparent recombinant shuffling of these variants among the seven ribosomal operons in each strain results in small clusters of SNPs and indels between ribosomal operons at comparable positions in whole-genome alignment of B and K.

The 5S rRNA genes are present in eight copies in each strain, twice in the D operon. The reference 5S sequence is found in six of the eight copies in B and in four of eight copies in K. SNPs are found in five different positions relative to the reference sequence, three in both B and K and one each uniquely in B or K. The second copy of 5S rRNA in the D operon of both strains contains the same set of three SNPs, and the remaining four variant copies each have one or two SNPs. In addition, two ribosomal operons of K have identical 104-bp replacements of 18 bp in a region not known to specify any RNA, and 3 ribosomal operons of B similarly have identical 147-bp replacements of 54 bp in a different region not known to specify any RNA. Each of the multi-base-pair polymorphisms and replacements in the ribosomal operons of B and K is found in at least one of the completely sequenced *E. coli* and *Shigella* genomes used for comparison, except for one variant in B that differed from its closest match by a single SNP. We do not know whether these relatively minor differences in the ribosomal operons have functional significance.

Deletions, insertions, and clusters in the basic genome

The most extensive interruptions of the aligned genomes of B and K are due to IS elements and large mobile elements, which will be discussed in later sections. However, aligned blocks are also interrupted at many sites by expansion or contraction of tandem repeats (usually in regions between coding sequences), by simple or complex deletions of various lengths, and by concentrated clusters of SNPs and indels that apparently can accumulate in nonessential regions or arise due to selective pressure. In this section, we consider only the interruptions that occur in well-aligned regions of the basic genome outside of large mobile elements and that are not caused by IS elements.

Simple deletions

Easiest to interpret are what we refer to as simple deletions, which remove a single block of sequence cleanly from one genome relative to the other. In aligning blocks of genome sequence, we have probably identified most simple deletions larger than ~50 bp, having cataloged 27 deletions of 62–8002 bp in MG1655 and 37 deletions of 52–10,021 bp in REL606. For simplicity, we refer to these interruptions as deletions, but at least 13 of them are due to expansion or contraction of tandemly repeated sequences of 92–483 bp located mostly between coding sequences. Slightly more than two-thirds of the simple deletions arose by crossovers at direct repeats at the ends of the deleted segment, with no crossover sequences apparent in the remainder. Thirty-seven of these deletions (55%) are in intervals between coding sequences, two of them are in ribosomal RNA regions, and one is in a region where other small RNAs have been identified. Fifteen simple deletions affect 27 annotated protein-coding sequences in REL606 and 10 deletions affect 28 coding sequences in MG1655, some coding sequences being completely deleted and others truncated or fused to other coding sequences. Simple deletions seem to have occurred more or less randomly in regions of the genome that are

dispensable under at least some conditions. Since both B and K have been maintained under various conditions in the laboratory for more than 80 years, it is quite possible that some of these deletions have arisen during that period and were not present in the original isolates.

Complex deletions

More perplexing are what we refer to as complex deletions, which leave unaligned sequences of different lengths between flanking, well-aligned blocks of sequence, giving the effect of having replaced a large segment of DNA in one genome with a smaller segment of unrelated DNA in the other. We analyzed an initial set of 22 complex deletions in the well-aligned basic genomes of B and K. These 22 deletions have a difference in the lengths of unmatched sequence in the two strains of 556–18,212 bp and a ratio of the longer to the shorter unmatched DNA of 4–183. In each case, the unmatched region is flanked by intact coding sequences in both strains (occasionally altered in a few codons at the C-terminus) and each deletion removed 1–15 coding sequences in one of the two strains: a total of 22 annotated coding sequences were removed by 8 complex deletions in B, and 64 annotated coding sequences were removed by 14 complex deletions in K.

The shorter unmatched fragment of these complex deletions has a recognizable remnant of a deleted gene in only one case, and attempts to identify the source of the small fragments present in the other complex deletions did not find a match in either genome. The elimination of almost all of the internal coding sequences without affecting the flanking genes indicates that selective pressures limited the extent of each deletion. The most likely explanation seems to be that a series of smaller deletions accumulated over time to the extent allowed without affecting the flanking essential genes, and the short unmatched sequences represent the residual DNA segments between multiple deletions, modified by accumulated SNPs. The few orfs of significant length not obviously related to a deleted gene are probably accidental in the remaining fragment. As noted previously, the *zjfZ* gene annotated in K is probably such an accidental orf. Support for this explanation of the possible origin of complex deletions comes from a subset of discontinuities in the genome alignments we refer to as clusters.

Clusters and complex deletions

We refer to a region of distinctly greater divergence within a larger block of well-aligned sequence of higher identity as a cluster. Clusters can easily be missed in large alignment blocks if not searched for specifically because they are often small enough to have little effect on the overall fractional identity. Typically, the flanking alignments will have >98% bp identity, whereas the embedded cluster of SNPs, or often SNPs and indels, covers a reasonably well-defined region from tens of base pairs to a few thousand base pairs, and the degree of divergence varies from moderate (~85–90% identity) to extreme (~50–75% identity). Although we have not completed a systematic search for and analysis of such clusters, approximately 50 have been identified in alignments of the basic genomes of REL606 and MG1655. Some of them seem likely to have arisen due to selective pressure, possibly coupled with horizontal transfer,[8,56] such as several clusters in sequences coding for proteins annotated as fimbrial-like adhesin or flagellar proteins and perhaps a few annotated as possible membrane or periplasmic proteins. Others lie entirely within noncoding intervals between protein-coding sequences.

Most of the clusters were of approximately equal length in the two genomes, as expected for diverging genes that remain functional in both strains, but a subset of about a dozen appeared to be smaller and more symmetric versions of complex deletions. Most were confined to noncoding intervals and none affected the flanking coding sequences. It seems likely that these represent early or intermediate stages in forming complex deletions by an accumulation of SNPs and deletions in a nonessential region. Further examination identified another five regions of unequal mismatch of tens to hundreds of base pairs and one region of several kilobase pairs that also have intact flanking coding sequences and seem likely also to be early-stage complex deletions.

Repeated horizontal transfer and homologous recombination appear to have integrated many DNA segments from diverged genomes into the genomes of B and K, as discussed in the earlier section on distribution of SNPs. It seems possible that complex deletions arose by the same process because each complex deletion is flanked by apparently essential genes that would provide matching DNA for genetic recombination even though the segments between them would lack any significant match. The integrated fragment that contained the complex deletion would presumably have come from a rather distant relative, where it would have had time to accumulate repeated deletions and SNPs in the nonessential region between the two essential genes.

Large mobile elements of B and K

Eleven large mobile elements or remnants thereof are annotated in MG1655, most of them thought to have integrated into the chromosome at specific sites through the action of phage-related integrases.[6,57,58] Seven of the same sites are occupied in REL606 plus four additional sites that are unoccupied in MG1655, giving a total of 15 insertion sites in the two strains (Table 7). Six sites are occupied by comparable elements in both strains, somewhat diverged from their common ancestral elements: DLP12, Rac, Qin, and KpLE2 are of comparable size and obviously related; the CP4-44 site contains highly diverged

remnants of a common ancestral element, as discussed in the subsection after next; and the P2 site of REL606 has a large, P2-related defective prophage where K strains have only a small remnant, as discussed in the earlier section on the P2*B defective prophage. The elements in REL606 that have no counterpart in MG1655 are named according to their attachment site followed by *B to indicate their host. P2*B and λ*B are found in known phage *att* sites and the other elements occupy core *att* sequences at or near the 3′ ends of genes that specify small RNAs, common sites of insertion for large mobile elements: *rybB* specifies a small RNA induced in stationary phase; *selC* specifies the tRNA for selenocysteine; *ssrA* specifies the tmRNA that participates in the rescue of stalled ribosomes; and *pheV* specifies a tRNA for phenylalanine.

The 24.2-kbp Rybb*B element is integrated at a 14-bp core *att* sequence that overlaps the 3′ end of *rybB*. This element is clearly phage related, with significant similarity to the P2-like prophages Fels-2 and SopEΦ, which, however, integrate into *ssrA*.[59] The *rybB* site is not occupied in MG1655, but some other completely sequenced *E. coli* strains have obviously related prophages at this site. Rybb*B carries retron Ec86, which specifies a reverse transcriptase and the unusual linked msRNA and msDNA, an element known to be present in B but not in K.[8]

The 20.9-kbp Selc*B element is integrated at a 17-bp core *att* sequence within *selC*, which was reconstituted upon insertion. Selc*B has signature features of CP4-like elements, discussed in the next section. The *selC* site is not occupied in MG1655.

The integration site for the 5.4-kbp Ssra*B element almost certainly overlaps *ssrA*, but the position of its core *att* sequence could not be determined directly because the right end of the insertion element has been lost due to an IS*1*-mediated deletion. However, the Ssra*B integrase has 62% amino acid identity to the somewhat shorter integrase of the Fels-2 prophage, whose 47-bp *att* core sequence overlaps the 3′ end of *ssrA*.[59] Furthermore, the coding sequence of each integrase begins the same distance from the nearest *att* site, and thus it seems possible that Ssra*B integrates at the same core *att* site, as Fels-2. Ssra*B is more than 99% identical with the first 5.4 kbp of a 34.0-kbp insertion in *E. coli* E24377A, which also seems to have lost its right end due to an IS*66* insertion. These elements carry plasmid partitioning genes, suggesting that they may derive from a large plasmid. The *ssrA* site in MG1655 is occupied by the apparently unrelated 22.0-kbp CP4-57, which appears to have an 8-bp core *att* sequence within *ssrA*, which is reconstituted upon insertion. The CP4-57 integrase has 48% amino acid identity to the Ssra*B integrase, and its coding sequence begins 37 bp further from *ssrA* than do the integrases of Ssra*B and Fels-2.

The 15.1-kbp Phev*B element of REL606 is integrated at a 17-bp core *att* sequence within *pheV*, which was reconstituted upon insertion. However, a deletion removed most of Phev*B from BL21(DE3). The *pheV* site is unoccupied in MG1655, but DNA

Table 7. Large mobile elements and their sites of insertion in REL606 and MG1655

Mobile element				Length of element	Length of *att* core L/R	Position in genome			
REL606	MG1655	Type	Insertion site			Left *att* core seq		Right *att* core seq	
att					51	267,390	267,440		
	CP4-6	CP4	*thrW* tRNA	34,308	51/51	262,122	262,172	296,430	296,480
DLP12			*argU* tRNA	24,539	47/47	536,869	536,915	561,408	561,454
	DLP12		*argU* tRNA	21,302	47/47	563,978	564,024	585,280	585,326
λ*B		?	λ	12,090	15/15	787,864	787,878	799,954	799,968
	att				15	806,551	806,565		
Rybb*B		P2	*rybB* RNA	24,155	14/14	880,541	880,528	904,696	904,683
	att				14	887,199	887,212		
att					29	1,211,617	1,211,645		
	e14		*icdA* C-term	15,204	29/29	1,195,471	1,195,499	1,210,675	1,210,703
Rac			*ttcA* N-term	17,496	26/26	1,409,094	1,409,069	1,426,590	1,426,565
	Rac		*ttcA* N-term	23,060	26/26	1,409,948	1,409,923	1,433,008	1,432,983
Qin			*ydfJ* N-term	29,897	11/11	1,600,054	1,600,044	1,629,951	1,629,941
	Qin		*ydfJ* N-term	20,469	11/11	1,630,311	1,630,301	1,650,780	1,650,770
CP4-44		CP4	Unknown	5131	?	Deleted	2,004,061	Deleted	2,009,191
	CP4-44	CP4	Unknown	12,688	?	Deleted	2,064,085	Deleted	2,076,772
P2*B		P2	*cyaR* RNA	22,146	23/23	2,100,285	2,100,307	2,122,431	2,122,453
	ogr-D′	P2	*cyaR* RNA	629	23/18	2,165,202	2,165,224	2,165,836	2,165,853
att					16	2,410,811	2,410,826		
	CPS-53 (KpLE1)		*argW* tRNA	10,215	16/16	2,464,391	2,464,406	2,474,606	2,474,621
att					8	2,486,210	2,486,217		
	CPZ-55		*eutA* C-term	6790	8/8	2,556,720	2,556,713	2,563,510	2,563,503
Ssra*B		?	*ssrA* (tmRNA)	5355	47/del	2,677,206	2,677,252	Deleted	2,682,607
	CP4-57	CP4	*ssrA* (tmRNA)	22,031	8/8	2,753,963	2,753,970	2,775,994	2,776,001
Phev*B		CP4	*pheV* tRNA	15,066	7/17	2,996,074	2,996,090	3,011,140	3,011,156
	att				17	3,108,441	3,108,457		
Selc*B		CP4	*selC* tRNA	20,893	17/17	3,775,335	3,775,351	3,796,228	3,796,244
	att				17	3,834,316	3,834,332		
KpLE2		CP4	*leuX* tRNA	36,681	14/del	4,479,698	4,479,711	Deleted	4,516,392
	KpLE2	CP4	*leuX* tRNA	39,771	14/del	4,494,493	4,494,506	Deleted	4,534,277

alignments show that Phev*B and CP4-44 of MG1655, about a megabase away, have similar genes and gene order over significant fractions of their lengths, shared also by genes of CP4-6 and CP4-57, as discussed in the next section. CP4-44 contains the *flu* gene of K, which specifies cell-surface antigen 43, whose phase-variable expression is controlled by *oxyR* and *dam*, and affects colony morphology.[60] The CP4-44 remnant in REL606 lacks a *flu* gene but Phev*B contains one. As with K, B strains that contain a *flu* gene can generate quite varied colony morphologies. Several B strains besides BL21(DE3) probably lack a *flu* gene because they failed to give a PCR product with primers designed to amplify the *flu* gene in Phev*B (Table 3). Those strains also seem to have a more uniform colony morphology, a trait that might have been selected in single-colony isolations. The *flu* genes of B and K are highly diverged, particularly in the center of their coding sequences, and represent two families of *flu* gene found in a variety of mobile elements similar to CP4-44 and Phev*B.

CP4-type elements

Three large mobile elements of MG1655 are annotated as cryptic P4 phages: the 34.3-kbp CP4-6, the 12.7-kbp CP4-44, and the 22.0-kbp CP4-57 elements; each contains a P4-type integrase at the left end and a cluster of similar genes or gene fragments at the right end, and CP4-57 contains *alpA*, which specifies a homolog of a P4 protein that can stimulate production of the integrase and cause excision of the element.[6,61] (We refer to the integrase end of integrated CP4-type elements as the left end, irrespective of orientation in the genome.) The lifestyle of the 11.6-kbp phage P4 provides an efficient way to mobilize DNA[62]: only a few P4 genes are needed for integration into or excision from the chromosome and to commandeer the structural apparatus of the 33.6-kbp helper phage P2 to package P4 DNA into phage particles for delivery to other bacterial cells; and P4 can also be maintained as a multicopy plasmid. Replacement of regions of P4 DNA not essential for mobilization can generate elements capable of using P2 to move unrelated DNA between bacterial cells. A wide range of P4-type integrases has been identified in sequenced genomes,[63] and it seems likely that different families of P4-like phages can be mobilized by helper phages with different DNA capacities. For consistency with previous literature, we refer to these mobile elements generically as CP4-type elements.

Phev*B appears to have been derived from an integrated CP4-type element by a deletion that truncated the integrase gene and fused it to a coding sequence ahead of a set of 15 right-end genes that seem to be mostly intact and characteristic of widespread families of CP4-type elements. The positions and lengths of the coding sequences in Phev*B are given in Table 8 along with the lengths of related coding sequences in Selc* B and in CP4-44, CP4-6, and CP4-57 of MG1655. The Phev*B genes are more closely related to those of Selc*B and CP4-44 than those of CP4-6 or CP4-57, which may represent a different family. The gene names are those previously given to genes in CP4-44 or CP4-6, plus a *cpf* designation for five genes of Phev*B not represented in MG1655.

Alignment of the Phev*B DNA sequence against whole-genome sequences of other *E. coli* and *Shigella* strains revealed many similar clusters of right-end genes, almost all of which have a gene or remnant annotated as a P4-type or phage-type integrase a variable distance to the left. Phev*B appears to retain a fairly complete representation of the genes characteristic of several families of CP4-type element, although its 570-bp *cpfE* gene is smaller than the more typical 846 bp in Selc*B and most other elements have only the C-terminal portion of *yeeU* (as do the other strains in the table). The genes most often found are *yeeP* and the nine genes *yafZ* through *cpfE*. As in CP4-6 and CP4-57, some families have additional genes between *yafZ* and *yafX*; others replace genes *flu* through *cpfC* with a different set of genes.

Although our sampling is not exhaustive, CP4-type elements occupy core *att* sequences in genes that specify small RNAs comparable to MG1655 genes *pheV*, *pheU*, *thrW*, *selC*, *leuX*, *serX*, and *ssrA*, as well as other sites where the core *att* sequence has not yet been identified. The boundaries of many elements are well defined by the presence of core *att* sequences at both ends, and lengths vary from a few thousand base pairs for highly reduced elements to well over 100 kbp, with most being several tens of kilobase pairs long. Although homology to genes of phage P4 has not been demonstrated for the right-end genes, their ubiquitous presence in these elements implies that they have important functions in mobilizing the associated DNAs and are probably homologs of genes in as yet unidentified families of P4-like phages. Most *E. coli* and *Shigella* genomes investigated have multiple copies of CP4-type elements (as many as seven per genome), and comparable elements are apparent in other species of bacteria. Clearly, CP4-type elements have been a major factor in moving DNA between bacteria and shaping their genomes. They appear to account for about half of the large mobile elements of B and K, with lysogenic phages from the λ and P2 families accounting for most of the rest.

The CP4-44 remnants of REL606 and MG1655

The relationship between REL606 and MG1655 in the interval between *cobU* and *yeeX* was difficult to decipher. The CP4-44 element in MG1655 has signature CP4-type genes at its right end but lacks an *int* gene or recognizable *att* sites, and the sequences of the two strains match only near the ends of the interval. More generally, genome sequences between *cobU* and *yeeX* are highly variable in *E. coli* and *Shigella* strains and typically contain many IS elements. However, *E. coli* E24377A

Table 8. Representation of genes typical of CP4-type elements in REL606 and MG1655

			REL606		MG1655		
Element			Phev*B	Selc*B	CP4-44	CP4-6	CP4-57
Size			15,066	20,893	12,688	34,308	22,031
att site			*pheV*	*selC*	?	*thrW*	*ssrA*
	Phev*B genes						
	REL606 position						
Genes	Start	End	Lengths of coding sequences				
attL	2,996,074	2,996,090	17	17	Deleted	51	8
int	2,996,294	2,997,253	960	1185	Deleted	1401	1242
Region occupied by carried DNA						17 orfs	9 orfs unrelated
cpfA	2,997,318	2,998,235	918				
yeeP	2,998,320	2,999,192	873		552	864	864
flu	2,999,564	3,002,410	2847		3120		
yeeR	3,002,531	3,005,047	2517		1533		
cpfB	3,005,124	3,005,579	456				
cpfC	3,005,658	3,005,891	234				
yafZ	3,005,991	3,006,809	819			822	822
						3 orfs	6 orfs related
yafX	3,006,864	3,007,349	486			459	459
yeeS	3,007,365	3,007,841	477		447	477	483
yeeT	3,007,904	3,008,125	222		222	222	201
yeeU	3,008,144	3,009,205	1062		369	318	318
yeeV	3,009,295	3,009,669	375	366	375	342	330
yeeW	3,009,666	3,010,154	489	489	168		
cpfD	3,010,166	3,010,363	198	243			
cpfE	3,010,460	3,011,029	570	846			
attR	3,011,140	3,011,156	17	17	Deleted	51	8

The right-end genes most characteristic of CP4-type elements are *yeeP* and *yafZ* through *cpfE*.
The Phev*B integrase gene is truncated by deletion after base pair 873.
Different families of *flu* protein are represented in Phev*B and CP4-44.
cpfD is annotated with either of two atg starts in different elements, for a coding sequence of 198 or 243 bp.
Most CP4-type elements have an N-terminal deletion in *yeeU*, relative to Phev*B.
A cpfE coding sequence of 843 or 846 bp is typical of most CP4-type elements that have this gene.

and *S. sonnei* are more than 98% identical across their ~17.5-kbp interval between *cobU* and *yeeX*, with no IS elements and only one small deletion in *S. sonnei*. Twenty-two genes occupy this region: an unspecified cobalamin synthesis gene adjacent to *cobU* and 21 genes involved in propanediol utilization (*pocR* and *pduA–pduV*). The cobalamin synthesis gene appears to be an N-terminal fragment of *cbiG*, a remnant of an ancestral cluster of genes for enzymes in the CobI pathway in cobalamin synthesis.[64] Seven different *E. coli* genomes, including REL606 and MG1655, aligned well with the genome of E24377A outside of the interval and lost alignment abruptly at the same base pair at each end of the interval (within the coding sequences for the two outside genes, the unknown cobalamin gene and *pduV*), thereby identifying the common endpoints for all of these insertions. The same endpoints, albeit complicated by nearby IS elements, are also evident in other strains of *E. coli* and *Shigella*.

The CP4-44 insertion in the *cobU–yeeX* interval of MG1655 is 12.7 kbp and that of REL606 is 5.1 kbp (Table 7), but the largest inserted sequences are those in CFT073, 536, and UT189, which are 45.0–47.3 kbp and align along their entire lengths with overall >99% identity and differ by only a few small deletions. Although these large insertions carry genes characteristic of CP4-type elements at their right ends, they lack the C-terminal portion of the *yeeW* gene and presumably the right end of the original element. Furthermore, they lack a P4-type integrase gene, indicating that the left end has also been lost. Conversion of the E24377A sequence in this region to that of the common ancestor of CFT073, 536, and UT189 may have been due to insertion of a CP4-type element larger than 47 kbp at a specific *att* site, which could have been anywhere within the deleted interval, followed by at least two deletions that removed both ends of the CP4-type element and the entire propanediol utilization gene cluster. The shorter, more varied DNA inserts in the *cobU–yeeX* interval of the other strains were apparently derived from this ancestral configuration by internal deletions, occasional insertions, and SNPs, since each insert has the same endpoints and aligns with 90% to >99% identity across much or all of its length with parts of the CFT073, 536, and UT189 sequences. In particular, REL606 has lost all of the characteristic CP4-type genes and its internal sequence aligns with a portion of the ancestral insertion that is not represented in MG1655, explaining why the remnants in the two strains align only near their ends.

KpLE2 is a remnant of a CP4-type element

The interval between *yjgB* and *yjhS* of B and K contains the KpLE2 mobile element integrated into the *leuX* gene, although the presumed *att* site at the

right end of KpLE2 has been lost[58] (Table 7). The KpLE2 elements of B and K are >99% identical except for regions differentially affected by IS elements, discussed in a later section. This site is highly variable in most of the other sequenced strains examined, containing remnants of CP4-type elements and multiple IS elements. Comparisons with four strains in this interval indicate that the KpLE2 elements of B and K are remnants of a CP4-type element that inserted into a 14-bp *att* core sequence corresponding to base pairs 66–79 of the 85-bp *leuX*, and which subsequently lost the right-end genes characteristic of CP-4 elements and a few adjacent genes ahead of *yjhS*. E24377A has a 55.3-kbp CP-4 element at this site that retains intact *att* core sequences at both ends and has most of the characteristic CP4-type right-end genes. Despite multiple differences caused by deletions and IS elements, an internal 10.5-kbp segment containing a cluster of seven *fec* genes and an adjacent highly diverged IS*1* element is >99% identical between the E24377A element and the KpLE2 elements of B and K. Likewise, *S. sonnei* has a 15.2-kbp element at this site containing a 9.1-kbp segment with >99% identity to the same *fec* genes and adjacent diverged IS*1* element. The *S. sonnei* element has lost the characteristic CP4-type right-end genes and *att* core sequence but retains an N-terminal fragment of the integrase coding sequence almost identical with that of KpLE2, with the same 191-bp spacing from the left-end *att* core sequence to the start codon (placing the start codon of the KpLE2 integrases of B and K ahead of an amber stop codon at codon 19). These integrases in turn are almost identical with integrases spaced 192 bp from the same *att* core sequence in 118.4-bp and 102.2-bp CP4-type elements in strains UT189 and 536. These two elements are closely related to each other, retain both of their *att* core sequences, and have characteristic right-end CP4-type genes, but they carry internal segments entirely different from those of KpLE2 and the E24377A and *S. sonnei* elements. Clearly, at least two distinct families of CP4-type element integrate at the *leuX* site in addition to phage P4 itself (which has a 20-bp *att* core sequence at the 3′ end of *leuX*). KpLE2 and the insertions in *leuX* of E24377A and *S. sonnei* are highly diverged from each other but apparently derive from the same ancestral CP4-type element.

Ribose-metabolizing genes specific to B

Both B and K have a cluster of six genes involved in ribose transport and metabolism located between *kup* and *hsrA* and designated *rbsD*, *rbsA*, *rbsC*, *rbsB*, *rbsK*, and *rbsR*. However, B has an additional cluster of genes comparable to the last five of these that interrupts a cluster of seven genes involved in fucose metabolism, located about a megabase away: a 6.8-kbp segment of DNA carrying genes comparable to *rbsA*, *rbsC*, *rbsB*, *rbsK*, and *rbsR* replaced 1.9 kbp of DNA between the start codons of the divergently expressed *fucA* and *fucI* genes, eliminating *fucP*. This appears to be a B-specific insertion, as none of the sequenced strains used for comparison has any type of insertion at this site. It is not clear to us how this apparently functionally redundant DNA was acquired, but the adjacent 219 bp in *fucA* and 15 bp in *fucI* are significantly less well matched to K sequence than is typical, reminiscent of insertion by mobile elements where genes interrupted by insertion are restored by fusion to sequences at the end of the insertion element. No integrase is apparent, but perhaps the mobilizing genes were deleted along with *rbsD*.

Rhs elements and ISEc1-type elements

K contains five regions, designated *RhsA* to *RhsE*, with common 3.7-kbp core sequences that code for part of relatively large proteins with variable extensions, other associated sequences, and a 1291-bp IS element that is variably inactive, truncated, or absent.[65] The *Rhs*-associated IS elements in MG1655, originally referred to as H-rpt, were renamed ISEc*1* to ISEc*6*.[66] B and K genome sequences are generally 98–99% identical in the *RhsA–RhsE* regions but also have deletions and mismatched regions. Their associated IS elements are quite variable, often fragmented, and none of them completely match a consensus sequence; we refer to them simply as ISEc*1*-type elements. B has a full-length ISEc*1*-type element between *yafT* and *yafV*, where K has only a 291-bp fragment, but neither strain has other obvious other signatures of an *Rhs* element. Unmatched sequences of several hundred base pairs in this region may represent different remnants of a common *Rhs* element previously at this site, and B has an IS*1* element that may have been responsible for deleting some of it. B also has a full-length ISEc*1*-type element between *nrfB* and *cusS*, embedded in 5.6-kbp of DNA not present in MG1655. The closest matches to this DNA in the nonredundant database have annotations, suggesting *Rhs*-related sequences. It seems likely that this DNA represents another *Rhs* region that remains, at least partly, in B but was eliminated from K by a 5.6-kbp complex deletion that left 26 bp of unmatched DNA in MG1655. As with other complex deletions, the flanking coding sequences remain intact, although the six C-terminal amino acids of *cusS* in REL 606 have been replaced by four amino acids in MG1655.

IS elements of B and K

PCR amplification and sequencing showed previously that most IS elements are inserted at different sites in B and K,[67] and prominent effects of insertion and IS-mediated deletions have been noted in the comparison of the complete genome sequences reported in the accompanying paper.[8] Detailed analysis of the genome sequences has identified all or parts of 12 different types of IS elements in 62 sites in REL606 and BL21(DE3), 54 of them in common between these two B strains (Table 9). Eleven of the same IS elements and two additional types have been identified in 54 sites in MG1655. However,

only 16 of the same sites are occupied in B and K, and thus most transpositions have occurred since the two lineages diverged from a common ancestor. The distribution of IS elements in the two strains suggests likely scenarios for their initial acquisition and subsequent transpositions, as discussed individually in the following subsections. ISEc*1*-type elements were discussed in the previous subsection.

IS1, *IS4*, *IS600*, and *IS911* were delivered by KpLE2

The CP4-type mobile element KpLE2 contains five full-length or partial IS elements at identical sites in B and K: from left to right they are IS*4*, IS*911*, IS*600*, a second IS*911*, and an IS*1* that is inactive and considerably diverged from the other IS*1*s of B and K. These five were probably acquired along with KpLE2 because they do not appear elsewhere in either genome except for IS*911* remnants in other large mobile elements. The single IS*4* is full-length in both B and K and differs by five SNPs from the reference sequence, taken to be the sequence of 17 identical copies in *S. sonnei*. The five SNPs change three amino acids in the transposase, but we do not know whether those changes are responsible for the lack of further transposition. The first IS*911* is full-length in B but inactive because of frameshifts in the coding sequence for the transposase, whereas the same IS*911* in K was truncated by insertion of IS*30* followed by a deletion. The deletion arose by a crossover between 8-bp direct repeats at the rightward end of IS*30* and internal to IS*600*, which is inserted into the second IS*911*. The IS*600* of B, on the other hand, is full-length and differs by six SNPs from the reference sequence, taken to be the sequence of a set of 11 identical copies in *S. sonnei*. The six SNPs change four amino acids in the transposase, but again, we do not know whether those changes are responsible for the lack of further transposition. Different segments to the left of IS*600* were deleted in B and K: the deletion between IS*30* and IS*600* in K removed end fragments from both of the IS*911*s and from IS*600* plus at least the *betU* gene between them; a deletion leftward from the full-length IS*600* in B removed the left end of the second IS*911* plus an indeterminate amount of DNA between it and the *betU* gene. The deletion in K was ~2.5 kbp greater than that in B.

The diverged IS*1*s in KpLE2 of B and K differ by 3 SNPs between B and K but by 72 or 73 SNPs from the IS*1* reference sequence, and both would be inactive because of internal stop codons in the frameshifted portion of the transposase. The reference sequence is taken to be that of nine IS*1*s of B and three of K: the other IS*1*s in B have 1–3 SNPs in five different combinations relative to the reference sequence, a pattern consistent with transposition and limited variation from a single founder that had the reference sequence. The likely founder is IS*1*-26, a second IS*1* in the KpLE2 of B, but not K, located 1.1 kbp to the right of the diverged IS*1* they have in common (IS*1*-25 in B). Differences between B and K in this region would be consistent with IS*1*-26 having arrived in the KpLE2 of both strains but having subsequently been deleted from the K lineage by a 3-bp crossover at the ends of the oppositely oriented IS*1*-25 and IS*1*-26. This deletion,

Table 9. IS elements in B and K strains

	REL606				BL21(DE3)				MG1655				Sites in common			
	Ref	Var	Frag	Tot	Ref	Var	Frag	Tot	Ref	Var	Frag	Tot	Ref	Var	Frag	Tot
IS*1*	8	20	—	28	9	20	—	29	3	4	—	7	—	1	—	1
IS*2*	—	—	2	2	—	—	2	2	6	—	1	7	—	—	1	1
IS*3*	2	3	2	7	2	2	2	6	4	1	2	7	1	—	—	1
IS*4*	1	—	—	1	1	—	—	1	1	—	—	1	1	—	—	1
IS*5*	—	—	—	—	—	—	—	—	9	2	—	11	—	—	—	—
IS*30*	—	—	1	1	—	—	1	1	2	1	1	4	—	—	—	—
IS*150*	4	—	1	5	4	—	1	5	1	—	—	1	1	—	—	1
IS*186*	5	—	—	5	5	—	—	5	3	—	—	3	2	—	—	2
IS*600*	1	—	—	1	1	—	—	1	—	—	1	1	—	—	1	1
IS*911*	—	1	2	3	—	—	3	3	—	—	3	3	—	—	2	2
ISEc*1*-type	—	4	3	7	—	4	3	7	—	3	4	7	—	2	4	6
ISEcB*1*	1	—	—	1	1	—	—	1	—	—	—	—	—	—	—	—
ISEhe*3*	—	—	1	1	—	—	1	1	—	—	1	1	—	—	—	—
ISZ′	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	—
Totals	22	28	12	62	23	26	13	62	29	12	13	54	5	3	8	16

The reference sequence is usually the majority sequence in MG1655.
The reference sequence for IS*4* is that of 17 of the 28 full-length 1426-bp elements in *S. sonnei*.
The reference sequence for IS*600* is that of 11 of the 51 full-length 1264-bp elements in *S. sonnei*.
The reference sequence for IS*911* is that of 5 of the 7 full-length 1249-bp elements in *S. sonnei*.
The reference sequence for ISEhe*3* is that of 11 of the 12 full-length 1229-bp elements in *E. coli* HS.
The reference sequence for ISEc*1*-type elements is the consensus sequence of full-length elements of B and K.
Variants are full-length or almost full-length with SNPs and/or small deletions, some obviously inactive.
A full-length element and a fragment at the same site are scored as a fragment site in common.
Two nearby fragments likely to have come from the same IS element are scored as one.

together with a 6.7-kbp rightward deletion caused by IS*1*-26 in the B lineage, would account for the observed nonmatching DNA between REL606 and MG1655 in the region. The IS*1*-26 equivalent of K would have transposed at least once before being deleted to account for the three IS*1*s of K having the reference sequence.

IS1, IS5, and IS30 were delivered to K by CP4-6

Three IS*1* elements of MG1655 differ from the reference sequence by nine SNPs. Two identical copies are in CP4-6 and an apparently transposed copy between *gfcA* and *cspH* has 7 of the same SNPs plus 2 additional SNPs. One or both of the copies in CP4-6 are probably the founder for these three IS*1*s. Thus, the seven copies of IS*1* in MG1655 seem likely to have arisen from founders delivered by both KpLE2 and CP4-6 (which is not present in B).

CP4-6 contains an intact IS*30* (inserted into a fragment of IS*911*) plus a fragment of IS*30* that has been truncated by an IS*1*-mediated deletion. Strikingly, the IS*30* elements inserted into IS*911* in CP4-6 and in KpLE2 are at exactly the same position in IS*911*, base pairs 334–335, but in the opposite orientation. A third intact IS*30* interrupts *ydbA*. IS*30* apparently was delivered by CP4-6 and transposed to both KpLE2 and *ydbA*, but the transposed elements have the same sequence as each other and differ by three SNPs from the intact IS*30* in CP4-6. Possibly the truncated IS*30* in CP4-6 gave rise to the transposed IS*30* elements before it was interrupted by IS*1*, or the SNPs were accumulated after transposition. The B lineage has only a single inactive fragment of IS*30*, in Selc*B.

CP4-6 is also apparently the source of IS*5*, which has 11 copies in K but is absent from B. The other large mobile elements containing IS*5* are unlikely to be the original source: the single IS*5*s in DLP12 and CP4-44 are much more likely to have transposed there from another site in K than to have arrived with them and subsequently been perfectly excised and lost from B, and the IS*5* in Rac is highly diverged, presumably inactive, and lies in a region that has been deleted in B. Nine of the other ten IS*5*s in MG1655, including the presumed founder in CP4-6, have the same sequence, and the tenth has five SNPs relative to the reference.

IS2

MG1655 has six intact copies of IS*2* and one fragment, all having the same sequence, whereas REL606 has only a fragment at the same site as an IS*2* in CP4-44 of MG1655 plus a highly diverged remnant in Selc*B. The unique history of the CP4-44 elements of B and K, discussed in an earlier section, together with the absence of an IS*2* element or remnant at the equivalent site in any other strain examined, argues that IS*2* was not acquired in CP4-44 but probably transposed to that site in the CP4-44 element of a common ancestor of B and K. This would imply that IS*2* was active in a common ancestor of B and K but that all active copies have been lost in the B lineage. Deletions induced by nearby IS*1* or IS*3* elements in B have removed regions where three intact IS*2* elements are found in K. An IS*2* in KpLE2 of K, but not B, probably transposed there after the B and K lineages diverged, as it is unlikely to have been perfectly excised in B. The origin of active IS*2* in the ancestor of B and K may have been Qin. This large mobile element has diverged considerably between the two strains but a fragment of IS*2* remains in the Qin of K opposite 2.2 kbp of unmatched sequence in B. It seems possible that an intact IS*2* was acquired in Qin and transposed in both B and K before being entirely deleted from the Qin of B and partially deleted from the Qin of K.

IS3

REL606 and MG1655 each have five complete or nearly complete copies of IS*3* but only one site in common, downstream of *ycdT*, which has an adjacent 1.7-kbp deletion in B but not K. This IS*3* is presumably the founder that populated the other sites in the B and K lineages by transposition, but it does not lie within a large mobile element. Perhaps it was acquired by transposition from a plasmid or some other mobile element that was subsequently lost.

hok genes, IS186, and IS150

B and K have five different *hok* genes in common, each of which specifies a 51-aa toxic protein that derives from *hok/sok* plasmid maintenance systems of conjugative plasmids, and B contains a sixth, designated *hokX*, that was lost from K by deletion, leaving upstream regulatory sequence.[68] The presence of multiple *hok/sok* modules and their remnants suggests that acquisition in the chromosome, presumably from plasmids, conferred a selective advantage, at least in some situations. However, these modules may be deleterious under other conditions, since the expression or activity of most of them is attenuated, usually by insertions of IS*186* or IS*150* but also by other mutations,[68] and several independent IS*150* insertions into *hokB* have been observed in evolution experiments growing populations founded by REL606 for many generations under well-defined conditions.[69,70] The distribution of IS*186* and IS*150* suggests that the founding members were acquired along with *hok/sok* modules.

The two sites of insertion of IS*186* in common between B and K are 22 bp and 21 bp downstream of the coding sequences of *hokC* and *hokE*, where they interrupted known regulatory elements of the *hok/sok* module.[68] These founder IS*186* elements apparently transposed once in the K lineage, inserting between *nupC* and *yfeA*, and three times in the B lineage, inserting 21 bp downstream of *hokX*, interrupting the *yeaR* coding sequence and causing UV sensitivity by interrupting the *lon* promoter, as discussed in the earlier section on UV sensitivity.

The IS*186* downstream of *hokX* in B caused a 9317-bp deletion that removed the eight *cas* genes present in K between *iap* and *cysH* (discussed in the next section in the subsection on 29–61 repeats, CRISPR, and *cas* genes). All of the IS*186* elements in B and K are intact and have the same DNA sequence.

The single site of insertion of IS*150* in common between B and K is just upstream of the coding sequence of *hokA*, where it interrupted regulatory elements for expression of *hokA* and caused a 39-bp deletion.[68] This IS*150* has remained dormant in K but transposed in B twice before divergence of the two sequenced B strains (upstream of *hokE* and in the coding sequence of *gltL*) and twice in each of the B strains after their divergence[8] (including inserting upstream of *hokX* in REL606). The IS*150* upstream of *hokE* apparently caused a 160-bp deletion and was subsequently interrupted by insertion of IS*1-7*, which also deleted 234 bp of this IS*150* element. Otherwise, the sequences of all of the IS*150* elements are identical.

ISZ′

An IS*4*-like element called ISZ′ has previously been described in MG1655 between *tdk* and *adhE*.[71] This 989-bp element is lacking in B but the 10-bp sequence CCCAGAAGGG, which was duplicated upon insertion in K, is present at the equivalent position in B.

ISEhe3 fragments

The equivalent fragment of ISEhe*3*, truncated after base pair 314 by IS*30* insertion and slightly diverged, is found in Selc*B of B and in CP4-6 of K. These two IS*30* elements (both of them terminal fragments) are inserted in the opposite orientation, as noted in an earlier subsection for IS*30* insertions into identical sites in IS*911*. Perhaps inversion of inserted IS*30* elements by recombination between their 12-bp terminal inverted repeats is not uncommon.

A new IS element, ISEcB1

A sequence in REL606 (2,332,096–2,333,536) and also in BL21(DE3), but not present in K, appears to be a new IS element, designated ISEcB*1*. This 1441-bp sequence is typical of IS elements: it ends in a 7-bp perfect inverted repeat that extends internally to 22 of 25 bp; it is flanked by a 4-bp direct repeat presumably generated upon insertion; and it codes for a protein of 154 amino acids that can extend to a 452-aa transposase-like protein by a −1 frameshift at the tandem lysine codons AAAAAG. The nucleotide sequence has no good match elsewhere in REL606 or K, nor indeed in the whole nonredundant database. The insertion site is to the right of the *yfbK*, *yfbL*, *yfbM*, *yfbN*, *yfbO*, *yfbP* cluster of genes found in MG1655, and all of those genes except *yfbM* are inactivated by a cluster of deletions and mismatches in REL606, including a 2544-bp complex deletion only 17 bp to the left of ISEcB*1*. We could not deduce a mechanism for involvement of ISEcB*1* in those deletions nor infer how it entered the genome of B.

IS-mediated deletions

Several IS elements have generated adjacent deletions of 0.2–40.8 kbp, especially in B, where IS*1* appears to have caused 10 deletions, IS*3* three, IS*150* one, and IS*186* one, for a total of 126.8 kbp deleted relative to comparable regions of K, affecting about 120 annotated orfs. Another IS*1*-associated deletion not apparent in these comparisons is the 6.0-kbp deletion responsible for the Mal$^-$ λ^R phenotype characteristic of B, which was repaired in the REL606 and BL21(DE3) lineages by P1 transductions from K, as discussed in an earlier section. IS elements have been much less active in causing deletions in K, with one deletion each apparently caused by IS*1* and IS*5* for a total of 13.2 bp deleted relative to comparable regions of B, affecting 14 annotated orfs. Differences in transposition and deletion activity may be due in part to differences in laboratory histories. Progenitors of the sequenced B strains are known to have been maintained for significant periods in sealed agar stabs,[14] conditions known to promote transposition.[72]

Notable differences between B and K due to deletions

In addition to the Mal$^-$ λ^R phenotype of the original B strain, deletions are responsible for several other characteristic differences between B and K. As already noted,[8] an IS*1*-associated 41-kbp deletion in B is responsible for its inability to methylate DNA cytosines and its lack of flagella and associated mobility. This 41-kbp deletion appears to be characteristic of B, since none of the dozen other sequenced *E. coli* or *Shigella* genomes examined has either IS*1* elements or a comparable deletion in this region, and PCR analyses confirmed that all of our B strains carry this deletion (Table 3). We here summarize other notable differences between B and K that are due to deletions of clusters of related genes or other interesting features, although these represent only a fraction of the genes affected by deletion.

Uptake of N-acetyl-galactosamine and galactosamine

E. coli B is known to be able to grow on *N*-acetyl-galactosamine and galactosamine, but K is not able to grow due to a deletion affecting *agaW*, *agaE*, *agaF*, *agaA*, which specify proteins necessary for the transport of these compounds.[73] A 2.3-kbp simple deletion in K relative to B is responsible for this difference.

Catabolism of aromatic acids and amines

E. coli B and K are known to differ in their ability to catabolize various aromatic compounds,[74] and

deletions are responsible for these differences as well. The B strains have the entire cluster of 11 *hpa* genes that specify enzymes for the catabolism of 4-hydroxyphenylacetate and related compounds, but this gene cluster has been completely eliminated from MG1655 by an 11.6-kbp complex deletion that left intact *yjiY* and *tsr* genes flanking 301 bp of unmatched DNA. On the other hand, MG1655 has a cluster of 17 genes (mostly annotated *paa*) that catalyze enzymes for the catabolism of phenylacetate and for oxidizing aromatic amines to acid substrates of this catabolic pathway. In B, an IS*3* insertion and adjacent IS-induced deletion removed all of the genes for phenylacetate catabolism and N-terminal portions of the flanking *tynA* and *ydbC* coding sequences.

Quorum sensing

MG1655 contains the *lsr* cluster of genes, which specify proteins crucial for quorum sensing through the autoinducer AI-2, thereby regulating expression of a range of different genes upon approach to saturation.[75,76] A 10.0-kbp simple deletion in B removed *lsrK*, *lsrR*, *lsrA*, *lsrC*, and part of *lsrD*, which specify the AI-2 kinase, transcriptional regulator, and three of the four proteins of the ABC transporter for uptake of AI-2. Therefore, B strains cannot respond to AI-2 in the medium, but we do not know how growth and regulation of B and K strains might differ as a consequence.

29–61 repeats, CRISPR, and cas genes

K contains two curious sets of 29-bp repeats that occur at 61-bp (or occasionally 62-bp) intervals, one set between *iap* and *ygbF* and the second between *ygcE* and *ygcF*.[77] Such repeats occur in B at the same sites, but with different numbers of repeats. The set following *iap* is truncated in B after the 5th of 13 complete repeats in K by an IS*186*-mediated deletion that also eliminated the eight *cas* between *iap* and *hokX*, ahead of *cysH*. In the region between *ygcE* and *ygcF*, 14 such repeats are found in B but only 7 in K. The 14 repeats in B plus flanking sequence could code for a protein of 282 amino acids, and reasonable upstream transcription and translation initiation signals are present. Because the 29-bp repeat occurs at intervals of 61 or 62 bp, the reading frame is different from one repeat to the next, so that motifs of 9 or 10 consecutive amino acids, usually separated by 11 nonrepeated amino acids, repeat 14 times in the potential protein. We do not know whether this protein is made or what its function might be, but many examples of predicted proteins that have matches to this pattern of motifs are present in the databases. Neither of the two sets of these repeats in K nor the other set of repeats in B has an orf across the entire set.

Given the typical near identity of comparable B and K genome sequences, it is remarkable that the 32-bp or 33-bp sequences that separate the 29-bp repeats have little similarity between B and K or even within or across sets of repeats in the same strain. Recent work provides an apparent explanation: K has been found to release ~57-nucleotide RNAs by specific cutting of transcripts made from the 29–61 repeat region that follows *iap* (referred to as CRISPR, for clusters of regularly interspaced short palindrome repeats).[78] These small RNAs interact with a protein complex referred to as Cascade, whose proteins are specified by some of the *cas* genes, to inhibit the growth of phages whose DNA contains sequences matched by the spacers. This mechanism is thought to provide a general defense against phage infection, but it is not clear how the matching spacer sequences might be acquired. Having lost the *cas* genes by IS*186*-mediated deletion, B strains apparently lack this phage defense mechanism but retain some of the associated repeated elements.

Protein secretion

Although K has a set of genes for type II secretion that can be made to function under special conditions, they are not expressed under typical laboratory conditions, and K has shown little capacity for protein secretion.[79] A progenitor of B and K apparently contained a second set of type II secretion genes in the region between *pheV* and *yghG*. These genes have been retained in B but an 8.0-kbp simple deletion removed *gspD*, *gspE*, *gspF*, *gspG*, *gspH*, *gspI*, *gspJ*, and *gspK* plus parts of the adjacent coding sequences from K.[8] This second cluster of type II secretion genes is present in most of the fully sequenced *E. coli* genomes and may be important for type II secretion. If so, B may have a capacity for protein secretion that is lacking in K.

Capsular genes

B contains a cluster of 14 *kps* and *kfi* genes that would specify a group 2 capsule of the K5 type except that they are disabled by an IS*1* insertion in the *kfiB* gene essential for polymerization of the capsule.[8,80] These genes lie between the core *attR* sequence of Phev*B and *yghD* in REL606 but have been entirely deleted from K by an 18.2-kbp complex deletion that left intact *pheV* and *yghD* flanking 100 bp of unmatched DNA.

Other deletions in B also prevent capsule formation. The *rcsA*, *rcsB*, *rcsC*, *rcsD* phosphorelay regulon activates synthesis of colanic acid extracellular polysaccharide and promotes group 1 capsule formation in response to certain environmental changes, and probably has a role in biofilm formation.[81] This system appears to be intact in K but *rcsA* is one of the genes deleted along with the flagella and DNA cytosine methylase genes by the 40.8-kbp IS*1*-associated deletion characteristic of B, and *rcsB* and *rcsD* have been eliminated by a 4.2-kbp IS*1*-associated deletion that also removed an N-terminal portion of *ompC*. The complete absence of *rcsA*, *rcsB*, and *rcsC* means that B should be unable to activate colanic acid synthesis and is presumably defective in biofilm formation.

Comparison of B and K with the Escherich strain

In tracing the lineage of B to a strain of the Institut Pasteur used by d'Herelle in 1918,[14] we speculated that perhaps this strain could have been directly descended from the first strain of *E. coli* described and cultured in the laboratory, by Theodor Escherich in Germany in 1885.[15] The Escherich strain was the type culture for *E. coli* in 1918 and was deposited in the UK NCTC in 1920 as NCTC 86. To test whether B might be a direct descendant, we obtained this strain from NCTC and sequenced small portions of its genome, initially in regions where we knew B and K to differ by as much as 5%. To our surprise, the Escherich strain was almost identical with K in the regions we initially sequenced, although K was known to have been isolated in Palo Alto, California in 1922.

Over the course of subsequently sequencing portions of the genome of different strains of B to sort out their relationships, we included the Escherich genome for comparison. We now have sequenced ~22.9 kbp of the Escherich genome in blocks of around 400 bp to 2800 bp distributed widely in the basic genome of B and K. We identified 339 SNPs between Escherich and B or K (1.4%), and at every SNP, two of the three strains have the same base pair. At these 339 SNPs, K has only 18 differences from the majority consensus, Escherich has 55 differences, and B has 266 differences. It appears from this limited sampling (~0.5% of the genome) that the basic genomes of the Escherich and K strains are more closely related to each other than either is to B. However, SNPs between B and K are distributed unevenly throughout their basic genomes, as discussed in the earlier section on distribution of SNPs, and a detailed analysis of the relationships among these three strains would require a complete genome sequence of the Escherich strain. In any case, it seems clear that these three laboratory strains were independently isolated and that B is not a descendant of the Escherich strain.

Overview and Summary

The availability of whole-genome sequences for both B and K allows a deeper understanding of the many studies of molecular genetics and bacterial physiology using these strains since at least the early 1940s, and provides specific explanations for known differences between B and K. From a practical standpoint, the ability to enumerate and understand every difference between the genomes of two B strains whose lineages are known and whose last common ancestor was at least 50 years ago illuminates the effects of some typical laboratory practices. We can now infer that the commonly applied MNNG mutagenic treatment causes ~50 unselected SNPs for each selected mutation and perhaps ~1–2 single-base-pair deletions as well. On the other hand, typical UV irradiation to a survival of 0.1% induces perhaps 8 deletions in the 0.2–4 kbp range and only ~2 SNPs. The ability of a significant fraction of the population not only to survive but also to maintain vigorous growth after such treatments testifies to the robustness of the genetic organization and regulatory network of *E. coli*.

The ancestral B strain had premature termination codons in two genes: *fnr*, a regulator that functions only under anaerobic conditions, and *btuB*, the vitamin B_{12} transporter, which is also a phage receptor. These mutations may have had little effect on strains maintained by streaking on agar slants made with nutrient broth, an early practice, but the active genes apparently conferred a selective advantage under other laboratory conditions because mutations that eliminated the premature termination codon arose independently at least four times for each of these two genes in different laboratory strains. Besides the MNNG- and UV-induced SNPs, no unselected SNPs were apparent between the two fully sequenced B strains, consistent with previous estimates of low rates of point mutation in *E. coli* with functional DNA repair and related pathways. However, two deletions of 5 bp, one of 12 bp, one of 18.1 kbp, a perfect tandem duplication of 82 bp, and a gene conversion between ribosomal operons occurred since the two lineages separated, indicating that such multi-base-pair events occur considerably more frequently under laboratory conditions than do SNPs. Three IS*1* and four IS*150* transpositions also occurred since separation.

The two B genomes show very different patterns of DNA integration in P1 transduction to Mal^+, which required restoring a 6-kbp deletion with DNA from K despite the B host restriction system. P1 could have delivered ~94 kbp of K DNA. In the BL21(DE3) lineage, little more than the deletion was replaced by K DNA, but in the REL606 lineage, at least six fragments of K DNA were incorporated over ~76 kbp of genome, a region that contains nine EcoB recognition sites and many SNPs (Fig. 1). These outcomes of P1 transduction illustrate the potential diversity of integration patterns in DNA transfers across restriction barriers in natural populations.

The basic genomes of B and K (the parts that remain after removing mobile elements) align typically with ~99% bp identity, interrupted by deletions and a few unmatched or poorly matched regions such as the gene clusters for O antigen or core LPS, which have been substituted by horizontal transfer.[8] The deletions appear to be of three types: (1) those caused by adjacent IS elements, (2) simple deletions that cleanly remove a segment of DNA with or without a crossover at direct repeats (or sometimes a contraction or expansion of tandem repeats), and (3) complex deletions. The first two types are easily understood, and these deletions may have been present in the original isolate from nature or may have occurred in the laboratory. Complex deletions, however, appear to have been produced by a progressive accumulation of SNPs and deletions of various sizes, leaving remnants of various lengths that usually have no obvious match in either genome. Almost invariably, complex

deletions are flanked by intact coding sequences that are presumably essential and cannot be eroded without serious consequence. Well advanced complex deletions probably have not had enough time to develop since the last common ancestor of B and K, and it seems likely that many or most of the complex deletions have been acquired by horizontal transfer from distantly related genomes and integrated into the B or K genome by homologous recombination in the conserved flanking genes.

Fourteen different types of IS elements or remnants thereof are found at 62 sites in each of the two B strains and at 54 sites in K, but only 16 sites are in common between B and K. In many cases, a good candidate for a founding IS element can be identified, most of which appear to have been acquired within larger mobile elements such as prophages. IS elements appear to have been more active in causing deletions in B than in K.

The large mobile elements in B and K also differ considerably. Particularly interesting is a remnant in B of what may be a P4-type prophage with an integrase at one end and a fairly intact signature of ~15 genes at the other end, the last 8 of which are particularly diagnostic. Related elements in K have been called CP4 (with an additional number that indicates position in the genome), and so we refer to them as CP4-type elements. Using the signature of this element from B, many related CP4-type elements can be identified in other sequenced *E. coli* genomes. At least eight different sites (mostly genes for small RNAs) are used for integration of different CP4-type elements, which can carry unrelated internal DNA segments of various lengths, extending to >100 kbp in some cases. As many as seven different CP4-type elements or remnants have been identified in a single genome. CP4-type elements appear to be a versatile family of mobile elements that can carry unrelated DNA of different sizes, and they have obviously been important in the evolution of *E. coli* genomes. About half of the large mobile elements of B and K appear to be related to the CP4 type, with most of the remainder appearing to be relatives of λ or P2.

A striking measure of the similarity of B and K is that more than half of the genes in their basic genomes (i.e., excluding mobile elements) produce identical proteins, and about a third of the coding sequences have identical base pair sequences. SNP-free genes are distributed around the genome in more than 500 regions separated by genes having a wide range of SNP density, with more than a hundred genes having 40 or more SNPs (Fig. 2). The distribution of SNPs around the genome appears to be consistent with inferences that the vast majority of SNPs in *E. coli* have been acquired by repeated horizontal transfer and homologous recombination, rather than by accumulation of single-base-pair mutations.[54,55] Whole-genome comparisons of the basic genomes of B and K with those of other, more recently isolated commensal strains of *E. coli*—currently underrepresented among sequenced *E. coli* genomes—may allow a more certain reconstruction of the ancestral genome of commensals, help elucidate evolutionary relationships, provide reliable estimates of divergence times and population structure of natural commensal *E. coli*, and perhaps identify typical changes that may occur in established laboratory strains relative to newly isolated ones.

Materials and Methods

Portions of the DNA of the strains listed in Table 1 were sequenced by PCR amplification, ABI dye terminator chemistry, and 3130*xl* capillary sequencer, using primers chosen from appropriate positions in the sequenced strains, and analyzed by Sequencher software (Gene Codes). Alignments and management of sequences and annotation for analyzing differences among complete genomes were done primarily with the Clone Manager program (Scientific & Educational Software) and Microsoft Excel spreadsheets, taking advantage of the initial automated and manual annotation.[8] Information about K genes[53] was accessed primarily through GenBank records and the EcoCyc Database,[82] and information about IS elements from the EcoGene[83] and IS Finder[84] databases. Automated comparison of the proteins of B and K was done using the program CompareGenes written in Python by P.D. and available from him.

Acknowledgements

We thank Eileen Matz and Mike Blewitt for technical assistance and the sequencing of different regions of the B and Escherich strains reported here, Chris Borland for first locating the *araA* mutation in REL606, and Haeyoung Jeong for preparation of Fig. 1. This work was supported by the GTL Program of the Office of Biological and Environmental Sciences of the U.S. Department of Energy and internal research funding from Brookhaven National Laboratory (F.W.S.); Consortium National de Recherche en Génomique (P.D.); the U.S. National Science Foundation and DARPA 'FunBio' Program (R.E.L.); contract DE-AC02-98CH10886, Division of Materials Science, U.S. Department of Energy (S.M.); and the 21C Frontier Microbial Genomics and Applications Center Program of the Korean Ministry of Education, Science and Technology, and the KRIBB Research Initiative Program (J.F.K.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.09.021

References

1. Gray, C. H. & Tatum, E. L. (1944). X-ray induced growth factor requirements in bacteria. *Proc. Natl Acad. Sci. USA*, **30**, 404–410.
2. Tatum, E. L. (1945). X-ray induced mutant strains of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **31**, 215–219.

3. Tatum, E. L. & Lederberg, J. (1947). Gene recombination in the bacterium *Escherichia coli*. *J. Bacteriol.* **53**, 673–684.
4. Delbrück, M. & Luria, S. E. (1942). Interference between bacterial viruses. I. Interference between two bacterial viruses acting upon the same host, and the mechanism of virus growth. *Arch. Biochem.* **1**, 111–141.
5. Demerec, M. & Fano, U. (1945). Bacteriophage-resistant mutants in *Escherichia coli*. *Genetics*, **30**, 119–136.
6. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
7. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S. *et al.* (2006). Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* **2**, 1–5.
8. Jeong, H., Barbe, V., Lee, C. H., Vallenet, D., Yu, D. S., Choi, S.-H. *et al.* (2009). Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 644–652.
9. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2000 generations. *Am. Nat.* **138**, 1315–1341.
10. Cooper, V. S. & Lenski, R. E. (2000). The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature*, **407**, 736–739.
11. Blount, Z. D., Borland, C. Z. & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **105**, 7899–7906.
12. Studier, F. W. & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130.
13. Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol.* **185**, 60–89.
14. Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S. & Kim, J. F. (2009). Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 634–643.
15. Escherich, T. (1885). Die darmbakterien des neugeborenen und säuglinge. *Fortschritte. der. Medizin.* **3**, 515–522.
16. Liao, D. (2000). Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J. Mol. Evol.* **51**, 305–317.
17. Hashimoto, J. G., Stevenson, B. S. & Schmidt, T. M. (2003). Rates and consequences of recombination between rRNA operons. *J. Bacteriol.* **185**, 966–972.
18. Arber, W. & Lataste-Dorolle, C. (1961). Erweiterung des wirtsbereiches des bakteriophagen λ auf *Escherichia coli* B. *Pathol. Microbiol.* **24**, 1012–1018.
19. Ronen, A. & Raanan-Ashkenazi, O. (1971). Temperature sensitivity of maltose utilization and lambda resistance in *Escherichia coli* B. *J. Bacteriol.* **106**, 791–796.
20. Dryden, D., Murray, N. & Rao, D. (2001). Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.* **29**, 3728–3741.
21. Studier, F. & Bandyopadhyay, P. (1988). Model for how type I restriction enzymes select cleavage sites in DNA. *Proc. Natl Acad. Sci. USA*, **85**, 4677–4681.
22. Lobocka, M. B., Rose, D. J., Plunkett, G., 3rd, Rusin, M., Samojedny, A., Lehnherr, H. *et al.* (2004). Genome of bacteriophage P1. *J. Bacteriol.* **186**, 7032–7068.
23. Cohen, D. (1959). A variant of phage P2 originating in *Escherichia coli*, strain B. *Virology*, **7**, 112–126.
24. Lederberg, S. (1966). Genetics of host-controlled restriction and modification of deoxyribonucleic acid in *Escherichia coli*. *J. Bacteriol.* **91**, 1029–1036.
25. Barreiro, V. & Haggård-Ljungquist, E. (1992). Attachment sites for bacteriophage P2 on the *Escherichia coli* chromosome: DNA sequences, localization on the physical map, and detection of a P2-like remnant in *E. coli* K-12 derivatives. *J. Bacteriol.* **174**, 4086–4093.
26. Rutberg, L. & Rutberg, B. (1964). On the expression of the *rII* mutation of T-even bacteriophages in *Escherichia coli* strain B. *Virology*, **22**, 280–283.
27. Slettan, A., Gebhardt, K., Kristiansen, E., Birkeland, N. K. & Lindqvist, B. H. (1992). *Escherichia coli* K-12 and B contain functional bacteriophage P2 *ogr* genes. *J. Bacteriol.* **174**, 4094–4100.
28. Witkin, E. M. (1946). Inherited differences in sensitivity to radiation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **32**, 59–68.
29. Witkin, E. M. (1947). Genetics of resistance to radiation in *Escherichia coli*. *Genetics*, **32**, 221–248.
30. Donch, J. & Greenberg, J. (1968). Ultraviolet sensitivity gene of *Escherichia coli* B. *J. Bacteriol.* **95**, 1555–1559.
31. SaiSree, L., Reddy, M. & Gowrishankar, J. (2001). IS*186* insertion at a hot spot in the *lon* promoter as a basis for Lon protease deficiency of *Escherichia coli* B: identification of a consensus target sequence for IS*186* transposition. *J. Bacteriol.* **183**, 6943–6946.
32. Johnson, B. F. & Greenberg, J. (1975). Mapping of *sul*, the suppressor of *lon* in *Escherichia coli*. *J. Bacteriol.* **122**, 570–574.
33. Mizusawa, S., Court, D. & Gottesman, S. (1983). Transcription of the *sulA* gene and repression by LexA. *J. Mol. Biol.* **171**, 337–343.
34. Huisman, O., D'Ari, R. & Gottesman, S. (1984). Cell-division control in *Escherichia coli*: specific induction of the SOS function SfiA protein is sufficient to block septation. *Proc. Natl Acad. Sci. USA*, **81**, 4490–4494.
35. Hershey, A. D. (1946). Mutation of bacteriophage with respect to type of plaque. *Genetics*, **31**, 620–640.
36. Unden, G., Becker, S., Bongaerts, J., Holighaus, G., Schirawski, J. & Six, S. (1995). O_2-sensing and O_2-dependent gene regulation in facultatively anaerobic bacteria. *Arch. Microbiol.* **164**, 81–90.
37. Chimento, D. P., Mohanty, A. K., Kadner, R. J. & Wiener, M. C. (2003). Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. *Nat. Struct. Biol.* **10**, 394–401.
38. Bassford, P. J., Jr. & Kadner, R. J. (1977). Genetic analysis of components involved in vitamin B_{12} uptake in *Escherichia coli*. *J. Bacteriol.* **132**, 796–805.
39. Bradbeer, C., Woodrow, M. L. & Khalifah, L. I. (1976). Transport of vitamin B_{12} in *Escherichia coli*: common receptor system for vitamin B_{12} and bacteriophage BF23 on the outer membrane of the cell envelope. *J. Bacteriol.* **125**, 1032–1039.
40. Studier, F. W. (2005). Protein production by auto-induction in high density shaking cultures. *Protein Expression Purif.* **41**, 207–234.
41. Funatsu, G. & Wittmann, H. (1972). Ribosomal proteins. 33. Location of amino-acid replacements in protein S12 isolated from *Escherichia coli* mutants resistant to streptomycin. *J. Mol. Biol.* **68**, 547–550.
42. Miller, J. H. (1983). Mutational specificity in bacteria. *Annu. Rev. Genet.* **17**, 215–238.
43. Kellenberger, E., Lark, K. G. & Bolle, A. (1962). Amino acid dependent control of DNA synthesis in bacteria

and vegetative phage. *Proc. Natl Acad. Sci. USA*, **48**, 1860–1868.
44. Lenski, R. E., Winkworth, C. L. & Riley, M. A. (2003). Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. *J. Mol. Evol.* **56**, 498–508.
45. Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA*, **88**, 7160–7164.
46. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, **148**, 1667–1686.
47. Ochman, H., Elwyn, S. & Moran, N. A. (1999). Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA*, **96**, 12638–12643.
48. Wood, W. B. (1966). Host specificity of DNA produced by *Escherichia coli*: bacterial mutations affecting the restriction and modification of DNA. *J. Mol. Biol.* **16**, 118–133.
49. Grodberg, J. & Dunn, J. J. (1988). *ompT* encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J. Bacteriol.* **170**, 1245–1253.
50. Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. (2001). Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* **183**, 2834–2841.
51. Echols, H. & Guarneros, G. (1983). Control of integration and excision. In (Hendrix, R. W., Roberts, J. W., Stahl, F. W. & Weisberg, R. A., eds), pp. 75–92, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
52. Weisberg, R. A. & Landy, A. (1983). Site-specific recombination in phage lambda. In (Hendrix, R. W., Roberts, J. W., Stahl, F. W. & Weisberg, R. A., eds), pp. 211–250, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
53. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R. *et al.* (2006). *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res.* **34**, 1–9.
54. Guttman, D. S. & Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, **266**, 1380–1383.
55. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P. *et al.* (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLos Genet.* **e1000344**, 5.
56. Milkman, R., Jaeger, E. & McBride, R. D. (2003). Molecular evolution of the *Escherichia coli chromosome*. VI. Two regions of high effective recombination. *Genetics*, **163**, 475–483.
57. Rudd, K. E. (1999). Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.* **150**, 653–664.
58. Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K. *et al.* (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157: H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**, 11–22.
59. Pelludat, C., Mirold, S. & Hardt, W. D. (2003). The SopEΦ phage integrates into the *ssrA* gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage. *J. Bacteriol.* **185**, 5182–5191.
60. Hasman, H., Schembri, M. A. & Klemm, P. (2000). Antigen 43 and type 1 fimbriae determine colony morphology of *Escherichia coli* K-12. *J. Bacteriol.* **182**, 1089–1095.
61. Kirby, J. E., Trempy, J. E. & Gottesman, S. (1994). Excision of a P4-like cryptic prophage leads to Alp protease expression in *Escherichia coli*. *J. Bacteriol.* **176**, 2068–2081.
62. Lindqvist, B. H., Deho, G. & Calendar, R. (1993). Mechanisms of genome propagation and helper exploitation by satellite phage P4. *Microbiol. Rev.* **57**, 683–702.
63. Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–875.
64. Lawrence, J. G. & Roth, J. R. (1996). Evolution of coenzyme B_{12} synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics*, **142**, 11–24.
65. Zhao, S., Sandt, C. H., Feulner, G., Vlazny, D. A., Gray, J. A. & Hill, C. W. (1993). *Rhs* elements of *Escherichia coli* K-12: complex composites of shared and unique components that have different evolutionary histories. *J. Bacteriol.* **175**, 2799–2808.
66. Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774.
67. Schneider, D., Duperchy, E., Depeyrot, J., Coursange, E., Lenski, R. & Blot, M. (2002). Genomic comparisons among *Escherichia coli* strains B, K-12, and O157: H7 using IS elements as molecular markers. *BMC Microbiol.* **2**, 18.
68. Pedersen, K. & Gerdes, K. (1999). Multiple *hok* genes on the chromosome of *Escherichia coli*. *Mol. Microbiol.* **32**, 1090–1102.
69. Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. (2000). Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, **156**, 477–488.
70. Woods, R., Schneider, D., Winkworth, C. L., Riley, M.A. & Lenski, R. E. (2006). Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **103**, 9107–9112.
71. Rudd, K. E. (1998). Linkage map of *Escherichia coli* K-12, edition 10: the physical map. *Microbiol. Mol. Biol. Rev.* **62**, 985–1019.
72. Naas, T., Blot, M., Fitch, W. M. & Arber, W. (1994). Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics*, **136**, 721–730.
73. Brinkkotter, A., Kloss, H., Alpert, C. & Lengeler, J. W. (2000). Pathways for the utilization of *N*-acetylgalactosamine and galactosamine in *Escherichia coli*. *Mol. Microbiol.* **37**, 125–135.
74. Diaz, E., Ferrandez, A., Prieto, M. A. & Garcia, J. L. (2001). Biodegradation of aromatic compounds by *Escherichia coli.Microbiol. Mol. Biol. Rev.* **65**, 523–569; table of contents.
75. Xavier, K. B. & Bassler, B. L. (2005). Regulation of uptake and processing of the quorum-sensing autoinducer AI-2 in *Escherichia coli*. *J. Bacteriol.* **187**, 238–248.
76. Li, J., Attila, C., Wang, L., Wood, T. K., Valdes, J. J. & Bentley, W. E. (2007). Quorum sensing in *Escherichia coli* is signaled by AI-2/LsrR: effects on small RNA and biofilm architecture. *J. Bacteriol.* **189**, 6011–6020.
77. Nakata, A., Amemura, M. & Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J. Bacteriol.* **171**, 3553–3556.
78. Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P. *et al.* (2008). Small

CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.
79. Francetic, O., Belin, D., Badaut, C. & Pugsley, A. P. (2000). Expression of the endogenous type II secretion pathway in *Escherichia coli* leads to chitinase secretion. *EMBO J.* **19**, 6697–6703.
80. Andreishcheva, E. N. & Vann, W. F. (2006). *Escherichia coli* BL21(DE3) chromosome contains a group II capsular gene cluster. *Gene*, **384**, 113–119.
81. Majdalani, N. & Gottesman, S. (2005). The Rcs phosphorelay: a complex signal transduction system. *Annu. Rev. Microbiol.* **59**, 379–405.
82. Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S. M. *et al.* (2007). Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.* **35**, 7577–7590.
83. Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.* **28**, 60–64.
84. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36.